Microsoft

Transforming Healthcare

Improving patient outcomes using the Microsoft Cortana Analytics Suite

Abstract

The healthcare industry is undergoing a technology upheaval as healthcare organizations transform their business models to comply with government mandates and work to improve operational efficiencies to drive down costs. The proliferation of healthcare data available—from new data sources streaming in real time to historical data stored on health provider systems—combined with the power of advanced analytics can help transform current healthcare business challenges into predictive and prescriptive solutions. This paper explains how the Microsoft technologies united within the Cortana Analytics Suite can help healthcare organizations quickly create solutions for their demanding and unique business needs.

Executive Summary

Success in the changing healthcare industry requires creating a data-driven organization for both high-quality patient care and profitability.

Preventable hospital readmissions in the U.S. account for an estimated $25 billion a year of wasted hospital spending and fines levied by the federal government. Healthcare organizations have a strong incentive to address this threat to their bottom line. Fortunately, recent advances in predictive analytics and cloud computing power are providing healthcare providers with the tools needed to protect revenue.

While a nurse, doctor or social worker may not be able to identify a patient that that will return to a hospital within 30-days, machine learning can. Classification algorithms such as support vector machines, regressions, and decision forests are capable of building complex models from vast amounts of historical data. These models can separate patients by high, medium and low risk. Once identified, medical personnel can intervene to prevent a patient's return to the hospital both during the stay and after discharge.

Developing and operating this solution requires enormous storage and many processor cycles. Before the cloud era, the cost of building this system would have made it impractical. Microsoft's Cortana Analytics suite is a game changer. The Cortana Analytics suite is an end-to-end Big Data and Advanced Analytics platform that provides everything necessary to build a readmission prevention system in one convenient package. System components are capable of acquiring data, processing it in real time, storing it and building models. The suite is also data scientist friendly and supports legacy industry standard analytics tools such as R and Python.

For the readmissions application, many components from the Cortana Analytics suite were used to build the solution. The following components were used:

- HDInsight – Hadoop Clusters
- HDInsight – Spark Clusters
- Azure Machine Learning
- Azure Steam Analytics
- Azure Data Factory
- Azure SQL Data Warehouse
- Azure Data Lake
- Power BI
- Azure Event Hubs
- Azure Web Applications

While the cloud provided the raw muscle for building the models, the user was provided with an interactive interface implemented in Power BI. Power BI is user friendly, low cost and makes it possible to create business value generating interfaces for the users quickly. Building a solution on a truncated timetable when billions of dollars are at risk from readmissions penalties can save millions.

Transforming Healthcare with the Cortana Analytics Suite

A big part of healthcare transformation in the U.S. is based on clinical and business decision-makers accessing data that helps them make decisions about patient care and limits the financial risk of Medicare/Medicaid fines. Such transformation requires employees in healthcare organizations to have rapid access to crucial information and insights from data across multiple relational and non-relational data sources. Existing systems describe what has happened, but in healthcare analytics, the ability to both predict what is going to happen and prescribe action to achieve a desired outcome can truly be the difference between losing a life and saving a life.

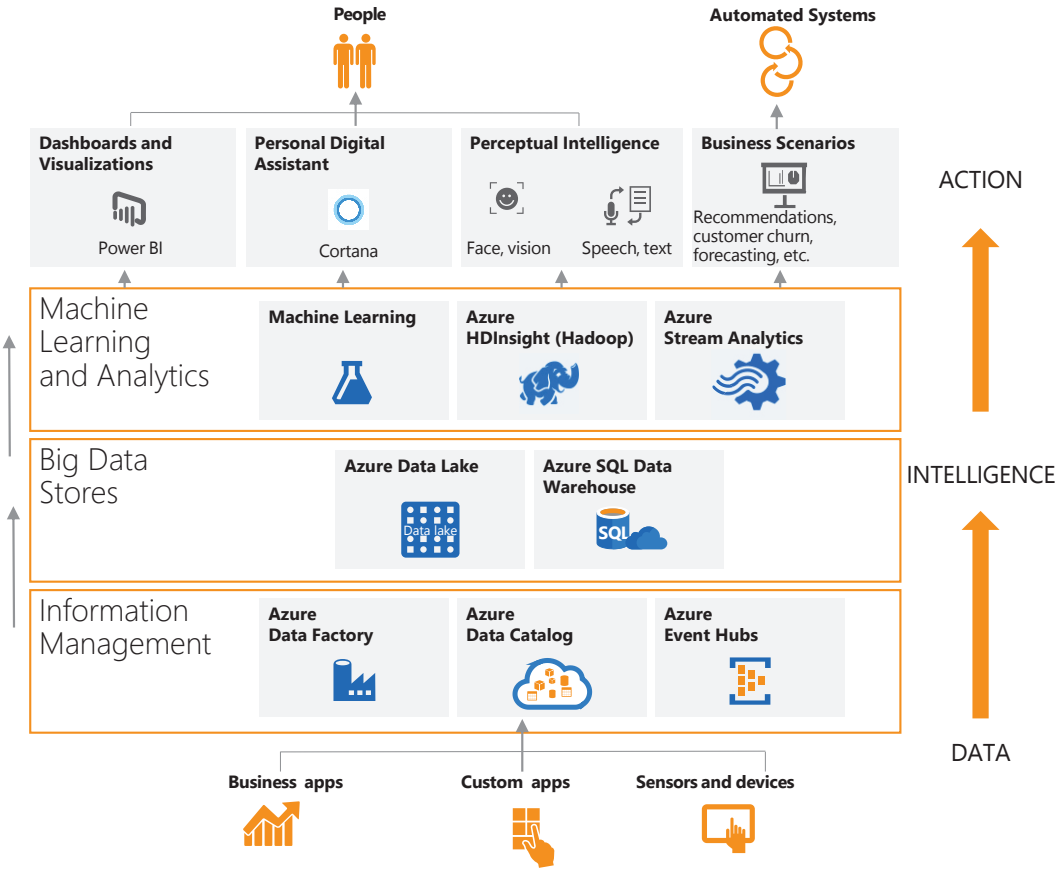
With the large amount of data potentially available for analysis, managing the flow of data efficiently can be a challenge. Rapidly feeding the results of data analysis to multiple devices is critical for a solution. An enterprise technology solution in healthcare transformation must solve the four "Vs" (volume, variety, velocity, veracity) of big data:

- Huge amounts of data to process (volume)
- A mixture of structured and unstructured data (variety)
- New data that's generated extremely frequently (velocity)
- Data quality so that it can be trusted (veracity)

Managing the four Vs is key to any healthcare solution. However, managing big data is a wasted investment if the data cannot be converted into intelligent action.

To achieve this, Microsoft's Cortana Analytics delivers a scalable data platform that handles large amounts of structured and unstructured data, whether its new data generated frequently or historical data residing in legacy database systems. With Cortana Analytics and its associated services, Microsoft securely **manages information** from multiple data sources—including data warehouses, healthcare monitors, sensors, healthcare apps, and outside sources such as other healthcare organizations and government databases. This data is catalogued, processed, and routed to **big data stores** for IoT operations or batch analysis. This data can then be used in **machine learning and analytics** to predict risks and outcomes, and the results can then be **visualized** through dashboards and services that run on multiple device types.

The following graphic illustrates the Microsoft technologies used at each step of the flow from data to action. Although an in-depth discussion of each service, product, and app is better suited elsewhere, you'll find a brief description of the technologies in the graphic as a reference. You can find links to more detailed information about individual products, services, and apps in the resources section at the end of this paper.



Microsoft provides the necessary tools to handle the large amounts of data flowing through a healthcare organization.

Building a Solution

The Affordable Care Act has created many new regulations for the healthcare industry. Non-compliance with these regulations means stiff fines for organizations, and those fines seriously affect the bottom line. The government cites patient readmission rates as one key area for improvement. One way that healthcare organizations can reduce the financial risk inherent in the new regulations is to focus on controlling patient readmission rates.

Big Data and advanced analytics can help healthcare organizations identify patients who are at risk for hospital readmission after receiving treatment. By combining real-time information about the patient's current stay in the hospital with historical information about the patient as well as external information, it's possible to identify a high-risk patient while he or she is still in the hospital. Care for the patient can be tailored in ways that prevent readmission—for example, using customized discharge planning, better coordination of post-treatment care, and planned patient follow up.

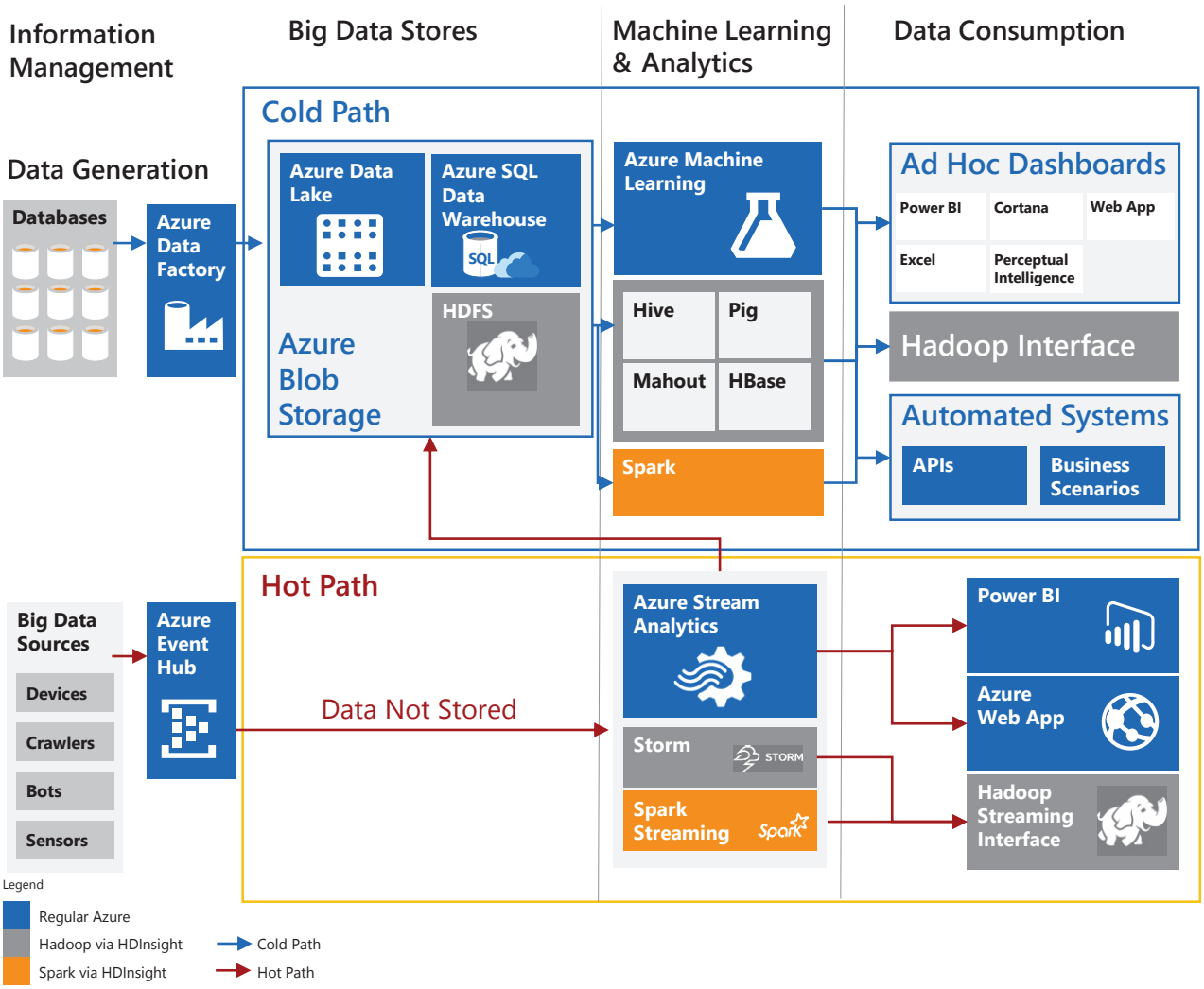
Data from sources including vital monitor, current Electronic Health Records (EHR), as well as visual records (such as MRIs, CT scans, ultrasounds, and x-rays) can be processed on the fly to detect and define risk indicators and send alerts about individual patients. Current Medicare/Medicaid regulations, billing information, and other government and/or insurance information can also be retrieved and combined with patient monitoring data to create a complete picture of the patient's risk of readmission.

With large quantities of historical data also available for analysis, such as clinical records and patient health monitoring over time, as well as external disease or condition statistics and data provided by the government and other healthcare organizations, a more complete picture of potential patient outcomes emerges.

Capturing this data is the first step in building an advanced analytics solution to meet organizational needs. However, that data is Big Data. Storing or processing such data in batch or real time requires both a scalable and flexible solution. To ensure privacy concerns are met, Personally Identifiable Information (PII) is obfuscated when necessary and stripped entirely when it isn't needed. For performance needs, sensor data must be filtered for relevance and all data must be transformed into the necessary format for analysis, including the creation of necessary calculated measures.

After the data is processed, it's ready to be used for analysis and predictions. Appropriate algorithms for analysis are selected, the patient outcome models are tested and fine-tuned, real-time alerts are defined, and a forecasting model can be built from the results of the analysis.

The end result of these steps is information that can be used both for real-time action and also forecasting purposes. The information can be visualized in a number of tools, such as patient advisor apps, alert apps and notifications, executive forecasting dashboards, and clinical reporting systems.



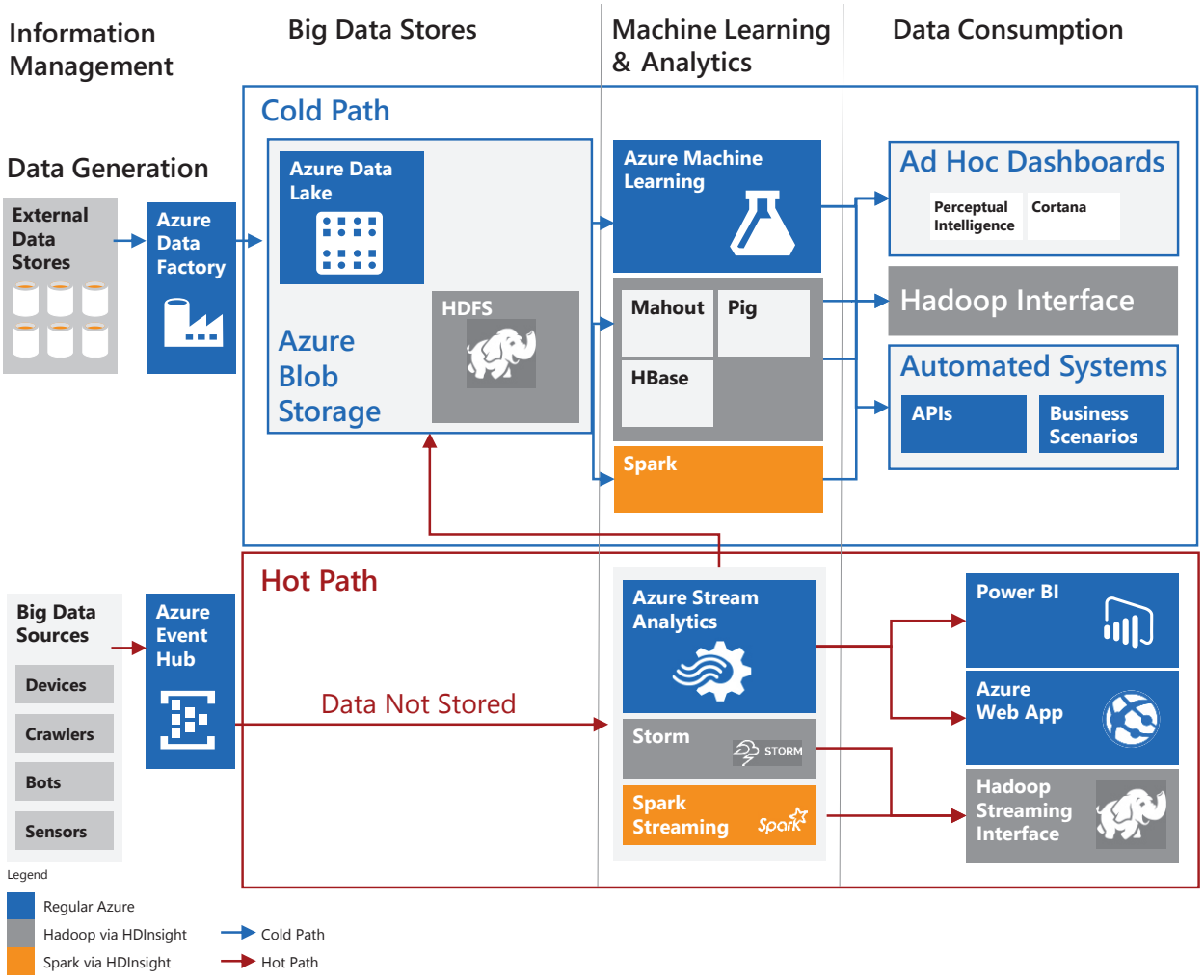
The capabilities of the Cortana Analytics platform can handle many possible solutions. One potential solution is explored in the next few sections.

Solution Overview

This section discusses one possible way to build a patient outcome risk solution using Microsoft products and services. This solution assumes that you have a modern EHR platform (such as EPIC's EpicCare or Cerner's PowerChart) to collect and manage patient EHR data, and medical sensors and devices that send data to the cloud via HTTP POST or AMQP 1.0, or that connects to a device that can.

The solution is still complex, but integrating these systems on Azure greatly simplifies the process.

Supporting infrastructure technologies, such as ExpressRoute, Azure Virtual Machines, Multi-factor Authentication, Azure Active Directory, and Azure Disaster Recovery are available to solve specific needs surrounding security, data redundancy, and access. As their configuration will be custom to each solution, they will not be discussed in this paper, but they are available and will likely be used in many solutions.



This diagram maps out the Cortana Analytics technologies in a solution built around managing patient outcomes.

The goal of the solution outlined here is to build a platform for managing patient outcomes that also allows executives to forecast the expected patient readmission rate. This requires the ability to capture, process, store, predict, and visualize data in real-time and also feed tools and apps that analyze the data over time.

This solution targets two primary user groups: physicians, (and other healthcare professionals in the hospital,) and executive-level decision-makers. The information generated by this solution also feeds back into the patient's EHR that's available to healthcare professionals the patient sees regularly outside the hospital. It can also feed into government data collection mechanisms and clinical research tools to create longer-term forecasting models for patient risk based on factors such as geography, time of year, demographics, and post–hospital-stay care.

Because the outcome of analysis is needed for both real-time and periodic analysis, it's necessary to create a solution for both:

- A real-time (also referred to as a hot path") solution that provides immediate feedback to healthcare professionals about the patient's status and potential risk factors allowing them to take quick action.
- A periodic (also referred to as a "cold path") solution that feeds into forecasting models.

The scenario discussed here uses Microsoft services and tools to build both types of solutions within the patient outcomes scenario.

Managing Patient Care

Aside from delivering the best care possible, the goal of managing patient outcomes is to increase the accuracy and timeliness of patient outcome predictions and warn of imminent critical conditions that may affect the patient both while in the hospital and also after discharge. This means collecting new data about patients during both patient intake and discharge surveys and also requires access to both real-time and historical patient information to generate an accurate prediction in time for it to be useful. During a patient's hospital stay, that could mean anything from improving the care itself by identifying complex interactions in sensors that indicate imminent critical conditions to educating patients on how to properly care for themselves when they get home. After the patient's hospital stay, it means providing personalized care solutions to patients flagged as having a high readmission risk, allowing further data collection and the ability to intervene before the patient ends up readmitted to the hospital. Personalized care solutions often involve follow up appointments with primary care physicians, phone calls to provide further education on care, or even connected devices used in the home to track vital statistics and monitor for abnormalities.

Information Management

For the hot path, collecting data from the myriad of sensors monitoring a patient's health can be done several ways. This solution uses the Event Hub service available within the Cortana Analytics suite to handle the ingestion of large volumes of sensor data. If the sensors can make requests via HTTP POST or AMQP 1.0, then they can connect directly to Event Hub. If not, an additional hardware investment is required, and while there are tools for this, it requires a custom application specific to the technology used by the hospital. For example, it's possible to use a Raspberry PI minicomputer to act as a gateway for the sensor to interface with Microsoft Azure services.

Streaming services with Microsoft Azure, like Azure Stream Analytics (ASA), Apache Storm on HDInsight, and/or Spark Streaming, provides real-time processing technology to run transformations and calculations on the data flowing in to process it before it's stored. This eliminates unneeded data and stores only the relevant data points. The data can be processed at essentially any interval, so that calculations update every second, every 10 seconds or 10 minutes, and so on. For the purpose of this paper, we'll consider Azure Stream Analytics (ASA) as the choice of streaming service used, owing to its simplicity and SQL-like interface.

Azure Stream Analytics connects directly to Event Hub for streaming data where very little configuration is needed. To set up an Azure Stream Analytics job, three components must be configured: inputs, query, and outputs. In this solution, the inputs are an array of event hubs collecting data from a set of identical sensors used by different patients. Within the hospital, health monitors are also wired to Azure, and are used to identify potentially dangerous conditions for patients.

The Azure Stream Analytics queries in this example are set up for two different functions based off of the data windowing within a query. To explain windowing, let's simplify to just examine heart rate. An alarm is set for the maximum heart rate exceeding 140, which is certainly an outlier for a bedridden patient. If a patient's heart rate spikes above that number, an alert timeframe that is short, such as 10 seconds is needed to quickly alert medical staff of the problem. However, because it's also useful to understand a patient's average heart rate above 100 over a longer period, say 10 minutes, a second Azure Stream Analytics job with a different query is used to track average heart rate over the longer window, because a prolonged elevated heart rate could be an indication of an imminent heart-related issue.

After the Azure Stream Analytics query processes the data, the output can be set so the results are visualized in real-time using Microsoft Power BI or by routing it back to Event Hub the data can be viewed on an Azure website or other internal tool. This data can also contribute to improving patient outcome forecasts, and to do so, the output can be set to store the data in Azure Blob Storage or a SQL Server database. Such a dataset can even be pulled into HDInsight clusters for further processing or used within the Machine Learning platform to track complex interactions and create a feedback loop for improving the stream processing queries.

If the scale of real time data is very large or more customizable functionality is desired, then there are big data technologies available within HDInsight to meet those needs. Spark Streaming and Storm are viable options to accomplish the same operations on streaming data and allow healthcare companies who are already vested and comfortable with big data to extend the value gained by adding an IoT capability to their solution portfolio. For this solution, Spark and Storm are used to relay IV fluid levels and consumption rate to ensure they are consumed at expected rates. With its close integration to the Hadoop and Spark ecosystems, these options are sometimes more favorable in terms of scale, extensibility and portability from existing on-premises to cloud solutions.

Regardless of the platform chosen, to minimize the volume of data processed in real time (lowering costs,) any solution should remove all data other than the actual sensor readings put it into a mapping table linked by device ID. That way things that don't change like device brand, function, room number, manufacture date, etc. aren't sent continuously with important data such as heart rate.

On the cold path, Azure Data Factory (ADF) is used to ensure data is moved through the various pipelines. An EHR system contains a vast amount of data that certainly qualifies as a big data problem. Storage itself is discussed below, but scheduling and orchestrating data flows is greatly simplified using

ADF. Azure Data Factory will extract data from the EHR platform and other sources programmatically, and feed the new data to the desired operation.

Processing data using Hadoop within an ADF job is very simple. An example used in this solution is a Hive query to generate the training dataset for a machine learning model to predict the likelihood of readmission based off of several behavioral markers and vital statistics. Once the Hive query is created, it is saved in the portal or can be passed to ADF via PowerShell using the JSONs that create the pipeline, source, and destination tables. Pig, MapReduce, and even SQL stored procedures can be orchestrated in the same manner. Simply upload the transformation action to be taken and set the scheduled execution interval.

Irrespective of the precise transformations needed, the HDI preprocessing generates a training data set for the machine learning platform, and with Azure Machine Learning, the batch processing or even retraining of the model can be scheduled to run as a part of the same pipeline. If an open source machine learning platform such as Spark ML or Mahout is in use, the dataset can still be prepped and piped into Azure Data Lake to be read by the analytics and ML platforms.

Big Data Stores

While data extracts from an EHR can be enormously valuable in their predictive power, the data is both very large and sensitive, so great care must be taken when interacting with it in the cloud. For this reason, removing personally identifiable information by anonymizing the data with an ID and mapping the patient ID back to a database that resides on premises is a secure solution that doesn't significantly impact performance. However, many of the machine learning scenarios in the healthcare vertical have no need for PII to be stored. This is because PII simply does not have any validity in a machine learning model. That is to say, the information linking Jon Doe to his medical history, such as his name, address, family doctor, or his parents' names, has no medical relevance and thus lacks predictive or prescriptive power. Data that is both sensitive and irrelevant should not be loaded into the cloud as a part of a solution as it is unnecessary.

Microsoft's premier option for storing limitless amounts of data in the cloud is Azure Data Lake, which has no individual file size or total storage consumption limits. Built for utmost throughput and extremely low-latency responses, Azure Data Lake is the ultimate store given its interoperability from structured querying layers along with the open source Hadoop and Spark ecosystems. Any amount of data to be used in predicting patient outcomes can be stored in the data lake and queried, read, or loaded into the variety of tools and platforms in the cloud or on premise systems.

For structured data, Azure SQL Data Warehouses can be used to host data so that it is readily available for querying by the variety of processing and prediction tools used in the platform. Having SQL available for simple bulk transformation and storage provides familiar flexibility while scaling to meet performance needs. Data workflows can be created to load data into Azure SQL Database using ADF or data can be loaded from on premise SQL Server deployments with relative ease.

Machine Learning and Analytics

An alarm triggered in Azure Stream Analytics is not a prediction per se, but it can be a useful tool to operationalize insights gained from machine learning experiments. Once a strong relationship is identified, it can be used to trigger alerts for medical staff to make them aware of changes in patient status that may not be immediately apparent otherwise. While updates to ASA queries must be configured manually, they comprise a valuable feedback loop between machine learning and the analytics tools that feed the models.

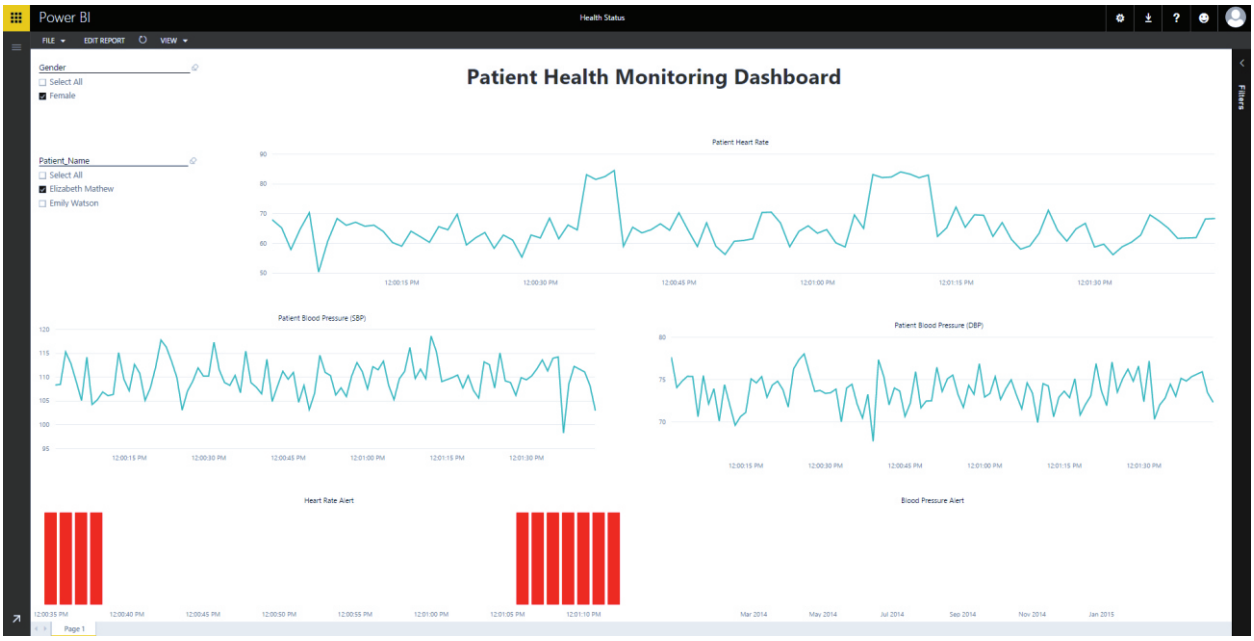
For true machine learning predictions in near real time, sensors can feed data to Azure Blob Stores, where Azure Data Factory both combines the sensor data with any other data sources and orchestrates a batch execution call to the Azure Machine Learning API. The call pings an ML model with the combined data and surfaces a set of predictions that an Azure Website, Power BI, or even Excel can display. Depending on the refresh interval and the data timeliness required, Azure Data Factory can be configured to process the data on a schedule that balances cost vs performance.

Visualize

There are many options to surface the insights provided by the analytics and ML tools. While the output of stream analytics can be shown directly on a Power BI dashboard, giving physicians and nurses the ability to check the vital readings of a patient from their tablet/computer anywhere in the world, the ability for Streaming tools such as ASA, Spark Streaming, or Storm to notify the appropriate parties in an emergency cannot be understated. When a dashboard incorporates the current vitals of a patient with the historical averages, recovery projections, and demographic data they can also compare their patient's vital signs to similar patients (anonymously, without PIIs) to see how the patient is doing relative to others who have undergone similar illness/surgery/treatment. The data refreshes automatically in the case of streaming data, but periodic data can be refreshed programmatically by ADF.

Excel can be used to make direct calls to an Azure Machine Learning web service, allowing for an interactive tool that is highly useful for the data science team to understand the relationships in the models they have built and if Excel is familiar territory, production class insight delivery can be built within Excel as needed.

Finally, because most of these projects are high value, high budget undertakings, a custom coded website integrated with Cortana Analytics is certainly possible and recommended for those looking for specific functionality not offered out of the box in Power BI. An example of such a website is included below, but Power BI offers an quick path to operationalizing sensor data streaming in from a variety of sources as shown here.



Power BI can display reports of individual patients as well as display alarms to quickly alert of critical issues.

Forecasting Patient Outcomes

While real-time data analysis provides critical information needed by healthcare professionals to provide up-to-the-minute patient care in a hospital setting, the use of the same data, combined with other historical information, enables healthcare executives to forecast patient outcomes, adjust patient intervention strategies, and appropriately staff and stock the hospital for improved patient care. Accurate forecasting is essential for complying with government regulations and evaluating financial risk factors in patient care.

The Cortana Analytics forecasting solution incorporates both the data from the hot path (sensors, current patient health at an aggregate level, etc.) that is diverted to the cold path and the existing cold path historical trend and contextual data. The data must be processed differently as the goal is not the same, resulting it filling a different portion of the data lake. Data duplication should be avoided but is not uncommon. However, joins will be common during the transformation process to get the perfect dataset for the forecasts. Predictions, once made, will be visible to a different class of end user entirely, so they will require a standalone visualization method. Typically these are targeted at executives, with different skillsets and backgrounds, so care must be taken in presenting the information at the right level for solid interpretation.

Healthcare information is captured and stored in EHR platforms that collect and manage many primary data points surrounding a patient's health. These internal systems log data from intake forms (including demographic data such as age, gender, weight and so on), doctor's notes, lab results, vital readings, and many more data points that paint a direct picture of patient status. This constitutes the first layer of context.

Using only internal data for predictions is limiting. Adding layers of situational or environmental data can be collected to improve data models. An example within the situational data layer could be the hospital's current risk of facing penalties. If the hospital is near the maximum number of readmissions allowed, they may be forced into a more conservative patient care strategy.

Environmental data that can be captured includes variables surrounding the current events in the region/hospital, such as an outbreak of MRSA or Flu season, which would indicate whether a patient should be released with a weakened immune system. It should also include any publicly available data from the government on patient activity, history, and region health.

For this scenario, a key component of potentially novel data collection within many hospitals is flagging whether a patient is readmitted for complications arising from an earlier visit that was covered under Medicare and subject to the restrictions and penalties surrounding this scenario.

Information Management

Assuming that the volume and velocity of the data is high, the primary method to process the data is HDInsight. HDInsight spins up Hadoop clusters on demand to preprocess and transform data as needed to prepare training datasets. Those datasets are then fed to Azure Machine Learning via ADF pipelines in the same manner as above. Even in a big data project, there are often smaller datasets that do not require Hadoop technology to process in a timely manner. For smaller datasets, SQL Server is often more than sufficient and can be spun up on an Azure VM as needed.

There are three primary goals during the Event Processing stage:

- Reduce the amount of data going in to the model to only the data relevant to answering the question for the scenario. In this case, we want only data that's useful in answering whether a patient is likely to return to the hospital unexpectedly or not.
- Create calculated measures and transform the data so that it's ready for modeling in the machine learning environment to generate the forecast. Hive, Spark, and other Hadoop transformations such as MapReduce are useful for this goal.
- Display the refined data used to generate the forecast in a tangible manner, usually a dashboard. This allows business decision makers to interact with the data and increase the trust in the forecasted results.

For example, the word count of certain key words in a field containing a physician's notes relating to the reason of the hospital visit can be useful, but because doctor's notes vary, they need to be standardized before they are used for modeling. Another example is creating a measure of the number of emergency calls for assistance to the medical staff. A count of such events creates a standard metric because comparing visits against each other leads to weighting issues. It's likely that dozens of calculated measures such as this would be created to ensure a fair and complete picture of patient health.

Azure Data Factory enables the hospital IT department to define a data-driven workflow on a processing cadence of their choice. As part of the data-driven workflow, Azure Data Factory can be used to invoke the published Azure Machine Learning web service APIs for prediction and retraining. Azure Data Factory can also be used to run data through Azure Machine Learning and feed the results back to Azure Blob Storage for further consumption.

Big Data Stores

Storage is handled similarly to the managing patient care scenario, so redundancy will be avoided. The key differences are as follows.

- All results for this insights pipeline are generated periodically and come from the cold path. Hot path data is used, but it comes from logs that have been stored in the Data Lake and further processed to add context to the raw sensor or machine data.
- Much more data is analyzed. While some predictive scenarios are likely to train off of datasets many terabytes in size, it is not uncommon for forecasting datasets to run off of many years' worth of data, often petabytes in size or larger. As a result, report frequency is significantly decreased as processing time increases.

Machine Learning and Analytics

Before data ever makes it to the Machine Learning models, it is imperative to prepare and pre-process the data coming in from a variety of sources and stores.

This is where the Azure HDInsight fits perfectly with the solution. With Hadoop/Spark tools and its scalability to process extreme amounts of data in a distributed fashion on multi-node clusters, HDInsight provides the perfect platform to address data quality concerns and form a core component of the agile Extract, Load and Transform (ELT) process as explained below:

Extract:

HDInsight provides scripting tools like Hive and Pig which assist in 'extracting' data from a variety of file formats such as XML, JSON, parquet etc. Even data compressed using widely used compression methodologies is easily processed. It is then transformed using a series of transformation routines. This transformation process is largely dictated by the data format of the output.

In this solution, the data being observed comes in a variety of formats like, patient data from relational systems, real-time data in the form of json, socio-economic data as xml and CDC or health records in flat files. All this data has a lot of garbage values that need to be dealt with and thus extracting data from these file formats and making them available for further processing is an essential task. We use Pig scripts to do these extractions and create a clean data to further load and transform.

Load:

Once the extraction is complete, we need to load this data in optimally performant file formats such that tools like Hive and Spark SQL can then carve informative data sets that allow users to use them for further processing.

In this case the format used is ORC since the tool of choice for transformation was Hive. But, this file format could have as easily been Parque, in case the Spark ecosystem was chosen to deploy.

Transform:

Tools like Hive and Pig in the Hadoop ecosystem and Spark SQL in the Spark environment provide easy Data quality and integrity checking is performed as part of the transformation process, and corrective actions are built into the process. Transformations and integrity checking are performed in the data staging area. Finally, once the data is in an acceptable stage, it is made available either directly to further layers in the solution and/or other tools within HDInsight for adding metadata to incorporate ease of use to the end user.

Azure Machine Learning is the keystone prediction service in Cortana Analytics. The goal is to predict whether a patient will return to the hospital with a complication arising from the current visit. Future visits that are unrelated must be classified as such. Spark ML, Mahout, and even custom coding a machine learning experiment are viable options depending on specific needs, but this solution used Azure Machine Learning as the ML platform.

Azure Machine Learning provides several efficient and effective classification algorithms that can be used to predict whether a customer will be re-admitted to the hospital due to complications. Fundamentally, whether a patient will readmit or not is a yes or no question, with only 2 possible answers. An extension of this scenario could involve sending the data where patients did return into a multi-class classifier to determine the leading contributor to determine precisely why the patient required a return visit, improving interventions and thus leading to more control over forecasted outcomes.

Once the patients are classified, the rate of readmissions can be calculated. Tools such as Logistic Regression even give weights to the various independent variables, so that healthcare managers can see what the most effective drivers of readmission are and develop interventions.

Two-Class Decision Forest/Jungle

A supervised ensemble learning classifier that uses a series of decision trees. It works by building multiple decision trees and then voting on the most popular output class. A decision jungle is similar but used when the dataset grows to the point that performance suffers from the decision tree having too many branches. A series of questions about the situation must be answered in order to determine the outcome. The machine learning algorithm will take the variety of data points surrounding a patient and

compare them to the trends among patients who are readmitted vs. those who are not readmitted.

A decision forest/jungle is most useful in a scenario when the logical boundaries surrounding patient classification are more clearly defined and can be well represented. For example, if the patient is known to be male, then all recommended interventions involving strictly female health considerations are ruled out. If they are female, then a series of questions can be answered to fill out a decision tree until the particular type of intervention best suited for the particular health issue is recommended.

Two-Class Logistic Regression

Logistic Regression is similar to linear regression in that it weights the independent variables' effect on a dependent variable. The key difference is that logistic regression is meant for a categorical dependent variable instead of continuous (linear). This works well to classify the dependent variable into one of two categories.

For healthcare providers desiring less of a black box model that is more managerially interpretable, logistic regression offers coefficients that data scientists can interpret as the driver for a particular action. This could be very useful for a hospital to diagnose the presence of gaps in their treatment process, resulting in an improved understanding of what drives positive patient outcomes at all levels of the business.

Two-Class Support Vector Machine

If there are relatively few data points (<100,000) or greater than 100 variables being used to compute the model outcome, a Support Vector Machine (SVM) can be more accurate/better performing than options A or B. An SVM is conceptually similar to a logistic regression in that it draws a line/wall through all the data points and splits them into two parts. One part is the first category, the other is the second category.

This means that SVM performs best on datasets where the interactions are more difficult to discern. For example, many factors could contribute to the readmission of a patient for a heart related condition, potentially hundreds. If the data is available, an SVM will discern the complex interactions and often provide a better prediction than other algorithms.

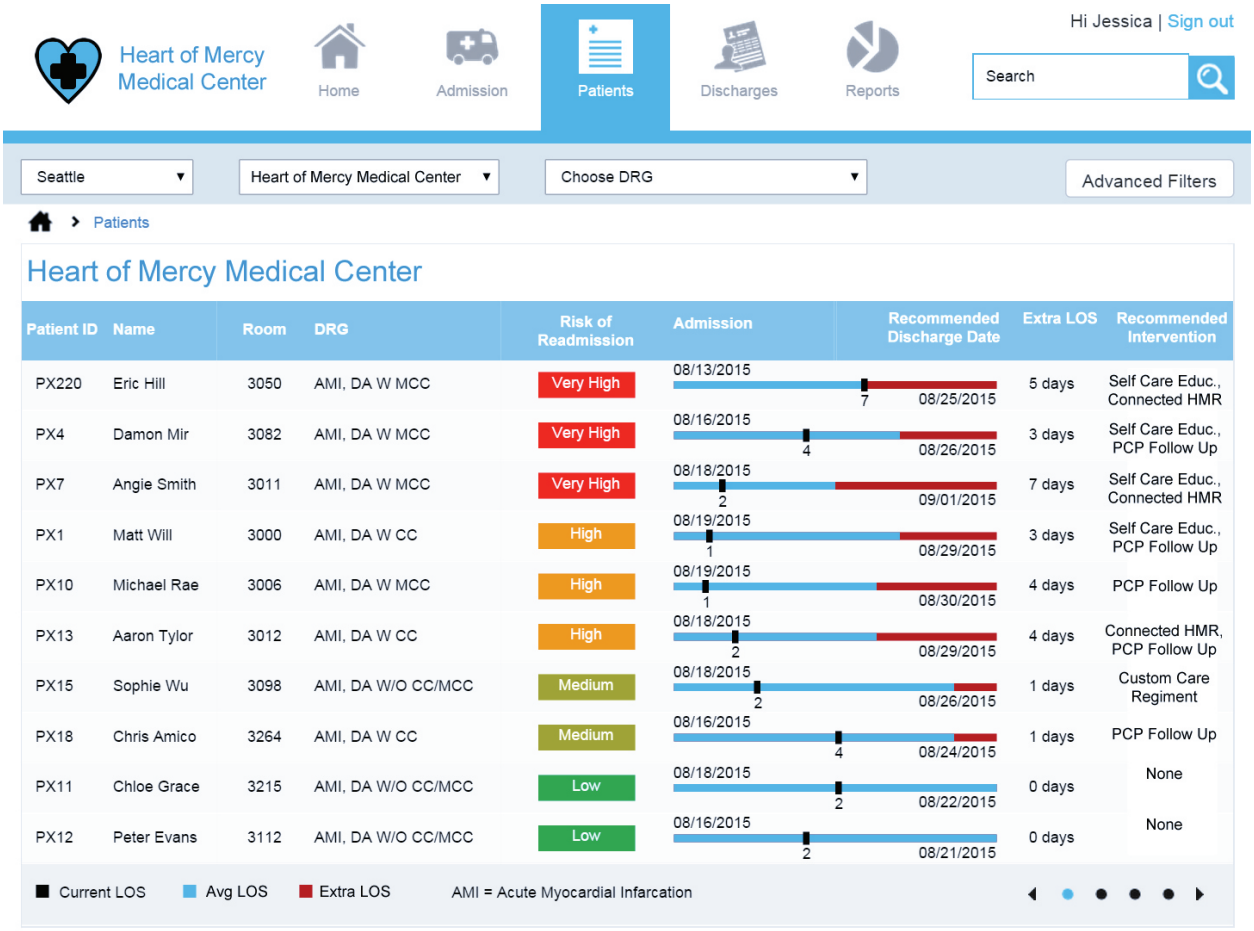
The selection of which prediction tool is dependent on each specific hospital's needs and data, but Azure Machine Learning provides options that fit the vast majority of use cases. If none of these work well, a custom Python or R-based algorithm can be developed to meet precise specifications/requirements or the other platforms mentioned above can be used.

After the system is implemented and interventions for high-risk patients begin, more cost-saving options become possible. A second prediction model can be used to recommend the best intervention method by maximizing the reduction in chance of readmission vs. the cost of the intervention.

Finally, a forecasting model can be built on top of all these to compare the normal expected readmission rate vs. the readmission rate if the recommended interventions are performed. Executives can then weigh the cost of fines vs. the cost of extra care.

Visualize

The results of the prediction process will likely be surfaced within the Hospital's existing systems, for example a web portal.



A custom Azure Web Application such as this mockup delivers exactly the desired information to the appropriate parties.

This web application could be accessed on tablets because it is a website, but it's also possible to build a Power BI dashboard to visualize the information on PCs and mobile devices. Aside from additional functionality, the key advantage to Power BI is that the time invested in building a workable dashboard is much less than a custom Azure Web App is, making it ideal for a phase one delivery method.

Conclusion

The Microsoft Cortana Analytics suite is a robust suite of tools and services that provides solutions for both big data and small data sets. For the healthcare industry, setting up a state of the art machine learning IoT solution to improve patient care and reduce readmissions can be accomplished quickly and with little upfront cost. Building a Big Data solution is a waste of resources without the proper framework to drive it to insights. There are plenty of tools in the market to choose from, however Microsoft Azure makes it easier and cheaper than the competition with an end to end solution in a single framework. It's how Dartmouth Hitchcock lowered costs while improving patient outcomes. It's how you can improve the care you provide patients while simultaneously lowering costs and avoiding penalties.

Appendix

Components of the Cortana Analytics Suite

The Cortana Analytics Suite encompasses a number of Microsoft tools and services to help efficiently process, store, and analyze data. With advanced analytics technology, such as perceptual intelligence and Cortana, the personal digital assistant, transforming data into action is easier than ever. The components of the suite include tools for information management, big data stores, and machine learning and analytics. The data can be fed into any number of dashboards and visualization tools.

Information Management

Azure Event Hubs

Azure Event Hubs ingest real-time data from sensors, devices, websites, and apps, and then process the data based on set rules. Azure analytics services can read the data provided by Azure Event Hubs, process the data, and feed the results back to Azure Event Hubs for ingestion by other services. To control access, permissions are assigned for writing to and reading from Azure Event Hubs.

Azure Data Factory

Azure Data Factory manages data movement and transformation workflows and can process on-premises data (for example, from SQL Server) with cloud data from Azure SQL Database blobs and tables. Workflows can be managed from a single tool, enabling you to set data production policy, identify and debug errors, connect, collect and compose data, and publish the data for analytics services. Azure Data Factory can even manage HDInsight tasks, moving and processing big data all the way through the intelligence pipeline.

Big Data Stores

Azure Data Lake

Azure Data Lake offers non-relational data storage, including blob, table, queue, and drive storage, and the data can be accessed programmatically. To secure the data, private keys to storage can be created and judiciously shared; for more restrictive access, access signatures can be required.

Azure SQL Data Warehouse

The Azure SQL Data Warehouse relational database service enables users to rapidly create, extend, and scale relational applications into the cloud. The data can be accessed programmatically using familiar SQL queries. As part of the scope of key Azure compliance certifications, Azure SQL Database has HIPAA approval.

Microsoft SQL Server

Microsoft SQL Server, Microsoft's long-trusted, robust enterprise database solution, also works well in a hybrid environment that spans on-premises and cloud computing and provides powerful business intelligence (BI) and other tools. New tools in SQL Server and Microsoft Azure make it even easier to provide an on-ramp to the cloud for organizations exploring hybrid cloud solutions.

Machine Learning and Analytics

Azure Stream Analytics

Azure Stream Analytics, an event processing engine, provides fast, real-time insights from devices, sensors, cloud infrastructure, existing data, and apps. Azure Stream Analytics integrates with Azure Event Hubs and can ingest and analyze millions of events to reveal patterns, detect anomalies, or kick off actions while data is being streamed in real time, all while using an intuitive, SQL-esque language.

Azure Machine Learning

Azure Machine Learning (AML) enables you to easily build, deploy, and share advanced analytics predictions and insights. AML helps you build and test predictive models, model data to predict future data, and enables more data driven insights and actions. AML also enables advanced analytics and helps map business problems to data, making it easier for data scientists at all skill levels to perform predictive analytics.

The Azure Machine Learning Gallery also provides access to sample APIs as well as the Machine Learning APIs published in the Azure Marketplace, enabling data scientists at any skill level to create breakthrough apps and perform sophisticated data analysis as well as try out advanced analytics features. The Azure Machine Learning Gallery community of developers and data scientists share solutions to interesting advanced analytics problems through the publishing of APIs and data experiments.

HDInsight

HDInsight is a cloud implementation on Microsoft Azure of the rapidly exanding Apache Hadoop technology stack that is the go-to solution for big data analysis. It includes implementations of Storm, HBase, Pig, Hive, Sqoop, Oozie, Ambari, and so on. HDInsight also integrates with business intelligence (BI) tools such as Excel, SQL Server Analysis Services, and SQL Server Reporting Services. Azure HDInsight deploys and provisions Apache Hadoop clusters in the cloud, providing a software framework designed to manage, analyze, and report on big data with high reliability and availability.

Query Workload: Hive, Pig etc.

While not the hot topic in Big Data discussions, much of the high value data for businesses is still structured and available in databases. To scale with the volumes of data available today, tools such as Hive and Pig can perform scripted actions and queries as needed to transform data for specific

applications. Hive can be utilized to mine structured data in a "SQL like" manner to aggregate and calculate KPIs for analysis on dashboards or to prepare data for machine learning algorithms. Pig can perform batch operations to quickly process data for other tools and workloads.

Storm

Storm is an established open source streaming data solution that cost effectively processes real time data feeds and outputs results at sub second latencies. With HDInsight, a storm cluster can be created and configured in minutes. Storm can consume data from many feeds using out of the box "spouts," or custom spouts can be written when needed. Data is processed and filtered such that only relevant and valuable information is stored or passed on to a real time dashboard.

Mahout

For an open source Machine Leaning platform, Apache Mahout runs natively on HDInsight for custom experimentation. Using the popular Data Science languages Scala and Java, Mahout offers many predefined algorithms for a wide variety of use cases that speed up the data science process by minimizing custom code.

Spark on HDInsight

Apache Spark is an open source processing framework that runs large-scale data analytics applications. Built on an in-memory compute engine, Spark is known for high performance querying on big data. It leverages a parallel data processing framework that persists data in-memory and disk if needed. This allows Spark to deliver both 100x faster speed and a common execution model to various tasks like ETL, batch, interactive queries, and others on data in HDFS. The Azure cloud makes Apache Spark easy and cost effective to deploy with no hardware to buy, no software to configure, a full notebook experience to author compelling narratives, and integration with third party BI tools.

Spark SQL

Spark SQL lets you query structured data inside Spark programs, using SQL like commands. Spark SQL in HDInsight is usable in Scala, Python and R. Spark SQL provide a common way to access a variety of data sources, including Hive, Avro, Parquet, ORC, JSON, and JDBC. You can even join data across these sources. Spark SQL reuses the Hive frontend and metastore, giving you full compatibility with existing Hive data, queries, and UDFs.

Spark Streaming

The streaming component allows for data collection from sensors and equipment so it can be analyzed in near real time. Similar to Storm, Spark operates by collecting data feeds and separates the valuable information from the noise. For example, when a sensor crosses a set threshold, an alarm can be pushed out to notify the appropriate party immediately, but there is little value in storing every second of the sensor data while it is operating within its normal range.

Spark MLlib

MLlib is Spark's scalable machine learning library consisting of common learning algorithms and utilities, including classification, regression, clustering, collaborative filtering, dimensionality reduction, as well as underlying optimization primitives.

Surfacing Actionable Intelligence

Microsoft Power BI

Microsoft Power BI collects data from across a variety of data sources, and presents the data to users, who can explore and visualize the data through a free-form drag-and-drop canvas, a broad range of modern data visualizations, and an easy-to-use report authoring system. Creating personalized dashboards within Microsoft Power BI allows at-a-glance reporting of defined data metrics, in real time and for periodic reporting needs.

Azure App Services

Whether on the web or mobile, with the Azure App Service, developers can rapidly build, deploy, and manage apps for their organizations using a single back-end. Developers can use their existing language skills, including .NET, Java, NodeJS, PHP, and Python. A rich gallery of pre-built APIs is also available via the Azure Marketplace. Azure App Services also provide enterprise-grade security and management to secure the data flowing to mobile apps.

Microsoft Excel

Microsoft Excel in Office 365 provides familiar tools, such as Power Pivot, Power View, Power BI, Power Map, and Power Query, which enable users to visualize and present data in meaningful ways. Excel provides business users with powerful analytics capabilities all within a familiar tool.

Resources

Data Factory
http://azure.microsoft.com/en-us/services/data-factory/

Data Catalog
http://azure.microsoft.com/en-us/services/data-catalog/

Event Hubs
http://azure.microsoft.com/en-us/services/event-hubs/

Data Lake
http://azure.microsoft.com/en-us/campaigns/data-lake/

SQL Data Warehouse
http://azure.microsoft.com/en-us/services/sql-data-warehouse/

Machine Learning
https://studio.azureml.net/

HDInsight
http://azure.microsoft.com/en-us/services/hdinsight/

Stream Analytics
http://azure.microsoft.com/en-us/services/stream-analytics/

Power BI
https://powerbi.microsoft.com/