The Value and Benefits of IBM InfoSphere BigInsights Running on IBM System z

IBM Redbooks Solution Guide

Information is power if you know how to extract value and insights out of it. The more that is known about a particular issue, situation, product, organization, or individual, the greater the likelihood of a better decision and business outcome.

Data is like oil because it can be refined and used in many different ways, increasing its market value. Unlike oil, however, it is a renewable resource.

Large enterprises have many applications, systems, and sources of data, some of which are used to fulfill specific business functions. Often, it is necessary and beneficial to bring these “islands” of information together to reveal a complete and accurate picture, ultimately getting closer to the truth by taking into account multiple perspectives.

The industry term that is used to describe the integration and analysis of multiple different sources of data (see Figure 1) to gather deeper insights from is known as *big data*.

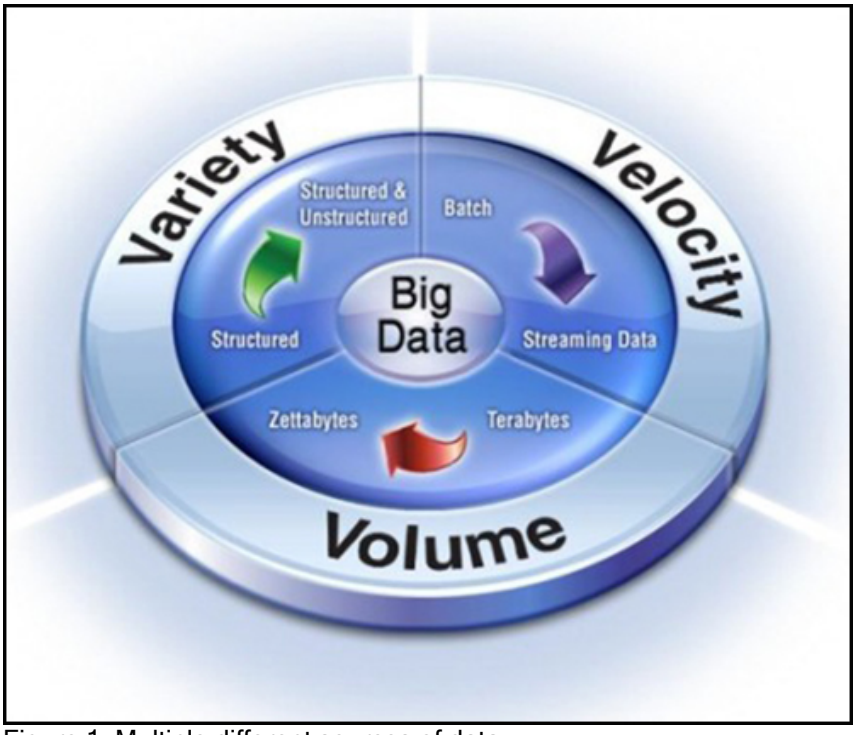

Figure 1. Multiple different sources of data

Did you know?

Today, most business analytics are based on information that is stored in enterprise data warehouses that are fed mainly from transaction and operational systems. This data is rich in value, is trusted and understood, as is its provenance. Used by 96 of the top 100 global banks, and 23 of the top 25 US retailers, IBM® System z® holds a significant amount of the world's business critical information.

Although valuable, this data on its own provides just one view of the world. The big data paradigm focuses on combining this data with many other information sources, such as social media, web logs, emails, documents, multi-media, text messages, and sensor information, providing a richer and complete view to augment our knowledge of the world around us.

New technologies, such as Hadoop, use a map/reduce paradigm that enables parallel processing of massive volumes of differently structured data that is spread across potentially hundreds and thousands of nodes. This breaks down the analysis of seemingly unmanageable data volumes into small discrete analytics jobs, and then the reduced result sets are combined to provide the complete answer. This IBM Redbooks® Solution Guide is intended to help organizations understand how IBM InfoSphere® BigInsights™ for Linux on System z and other related technologies can help deliver improved business outcomes as part of a big data strategy.

Business value

The interest and uptake of Apache Hadoop in the market has been described as unstoppable by analysts, including Forrester Research. The appeal of no-cost open source software and low-cost commodity hardware favors a "divide and conquer" parallel processing approach to analyzing large semi-structured and non-structured data sets. But what starts as an experiment of low cost, "good enough" hardware and software often falls apart when this situation is applied to mission-critical data. A challenge that clients face in big data initiatives is efficiently extracting, transforming, and loading (ETL) large volumes of data from sources such as IBM DB2®, IBM IMS™, and VSAM into Hadoop clusters in a timely and cost-efficient manner.

One critical decision clients make is choosing where to analyze the data. This decision is often influenced by where the data originates and the classification of the data's sensitivity.

IBM InfoSphere BigInsights elevates "good enough" Hadoop to an enterprise-ready, business-critical analytics solution. IBM InfoSphere BigInsights for Linux on System z, combined with IBM InfoSphere System z Connector for Hadoop, provides customers with two key advantages:

- Data that is stored on System z can remain on the platform for analysis under the security of System z while efficiently and effectively moving DB2, IMS, and VSAM related data from IBM z/OS® to Hadoop clusters on Linux on System z partitions or off platform clusters.
- Organizations with sensitive information can ensure that it remains secure by keeping it on System z or moving it there from less secure environments.

The combination of IBM InfoSphere BigInsights for Linux on System z and IBM InfoSphere System z Connector for Hadoop represents lower business risk by reducing the potential of a data breach, lower costs of managing and moving the data between traditional and Hadoop environments, and provides opportunities for growth by offering deeper analysis and insights of data that is stored or needed by applications on System z.

Solution overview

IBM InfoSphere BigInsights on the System z mainframe (mainframe) essentially allows clients to have the best of both worlds. They can continue to benefit from the security and reliability of the mainframe for processing critical data, but they can simultaneously take advantage of the rich tools that exist in Hadoop without compromising the security of operational systems. By combining mainframe data with data from other sources, organizations can obtain a complete view of their business and often gain insights that can help them improve efficiencies, find new revenue opportunities, or reduce costs. The following sections describe how Hadoop is deployed on the mainframe from an architectural standpoint, the software capabilities in InfoSphere BigInsights, and some of the unique considerations that must be taken into account when running Hadoop on the mainframe.

Solution architecture

IBM InfoSphere BigInsights is an industry-standard Hadoop offering that combines the best of open source software with enterprise-grade capabilities. BigInsights can simplify deployment and accelerate time to value for various big data and analytic workloads both on and off the mainframe.

System z customers can run BigInsights natively on System z Linux partitions, or external to the mainframe on system nodes that are connected through 10 GbE, as shown in Figure 2.

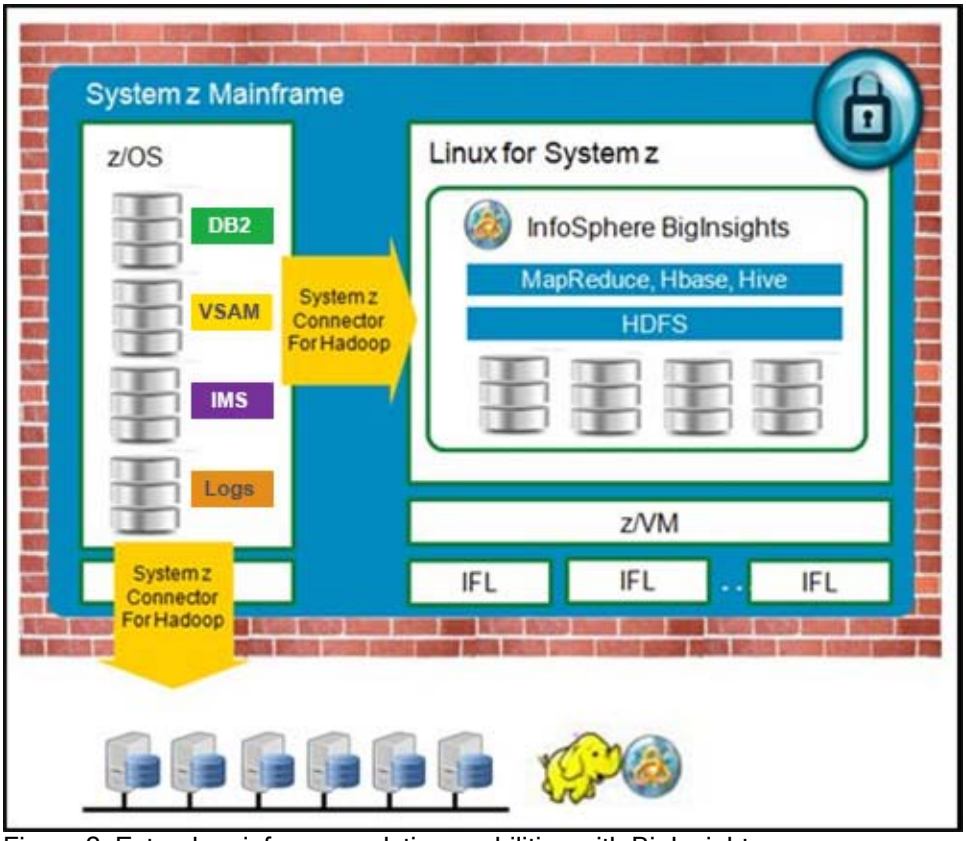

Figure 2. Extend mainframe analytic capabilities with BigInsights

There might be significant advantages to running on the mainframe depending on customer requirements:

- Hadoop applications can exist within the System z security perimeter.
- Clients can use mainframe technologies, including IBM HiperSockets™, to securely access production data, and move that data to and from Hadoop for processing.
- Clients can realize the management advantages of running Hadoop on a private cloud infrastructure, providing configuration flexibility and virtualized storage, and avoiding need to deploy and manage discrete cluster nodes and a separate network infrastructure.
- Clients can extend System z governance to hybrid Hadoop implementations.

Clients might prefer to run BigInsights on external cluster nodes that are connected to the mainframe. This might be the appropriate decision in cases where there is less concern about the security of the data, or where data volumes are large but non-critical to the business. This still allows for the results sets to be uploaded into a DB2 for z/OS based data warehouse to, for example, augment a customer record.

IBM InfoSphere BigInsights runs on Intel systems from various manufacturers or on IBM Power Systems™ for clients that need a higher level of performance and reliability than is offered by commodity systems.

IBM InfoSphere BigInsights' software architecture is platform-neutral; it remains consistent whether Hadoop is deployed on or off the mainframe. The following sections look at considerations that are specific to deploying BigInsights in a System z environment.

IBM InfoSphere BigInsights software capabilities

This section looks at some of the software capabilities of InfoSphere BigInsights and describes how it is implemented in the System z environment.

A primer on Hadoop

At a high level, Hadoop is a distributed file system and data processing engine that handles high volumes of data.

Hadoop was inspired by Google's work on its distributed file system (called GFS) and the MapReduce programming model. Google was wrestling with data volumes that at the time were unprecedented. Existing commercial data management systems could not cost-efficiently address Google's business requirements.

Early work that led to Hadoop was conducted by developers of Lucene (a search indexer) and Nutch (a web crawler). Douglas Cutting, who worked at Yahoo at the time, provided leadership on both of these projects by supporting Yahoo's internal search implementation. Today, Douglas Cutting is credited with the development of Hadoop and its transfer to the public domain as a top-level Apache project.

Hadoop has continued to evolve. Today, there are many subprojects and related projects in the Hadoop ecosystem.

Hadoop uses a different architectural approach than other data management systems. The heart of Hadoop is the Hadoop Distributed File System (HDFS). HDFS is designed for massive scalability across thousands of commodity computing nodes and tens of thousands of physical disk drives. A key design philosophy behind Hadoop is "design for failure" because at this scale, nodes, disks, networks, and operating systems certainly fail. Hadoop is designed to tolerate failures and uses software techniques such as block-replication and synchronization and service failover to ensure continuous availability of data and cluster services.

Other key Hadoop design principles are parallelism and the notion that it is more cost-efficient to move execution logic to data rather than move large data blocks across congested networks. The MapReduce algorithm reflects this design principle. Mappers are dispatched to run in parallel across a distributed cluster to process vast data sets quickly. Because map tasks are simply programs (usually written in Java), data can exist in virtually any format. Although Hadoop is used regularly to process structured or semi-structured data, it also can be used to process unstructured data if mappers can parse the contents of data blocks in HDFS.

Corporate IT organizations, increasingly faced with demands to retain and process vast amounts of data, have recognized the importance of Hadoop. Many organizations, including IBM, are providing Hadoop toolsets for corporate computing users to solve various problems that are difficult or costly to solve by using traditional techniques.

Hadoop on the mainframe

Hadoop evolved in commodity environments, so the implementation of Hadoop in a highly virtualized environment like the System z mainframe poses some challenges:

- Cluster nodes: In Hadoop terminology, a cluster node refers to a physical server that is connected to a Internet Protocol network. Hadoop servers are typically “dense”, that is, they are composed of many processor cores, much memory, and multiple hard disk drives (HDDs). A typical configuration for a commodity Hadoop “node” might be a system with two six-core processors (for a total of 12 cores) and 12 HDDs along with 128 GB of memory. In the context of the mainframe environment, a Hadoop “node” is typically a virtual machine running the Linux operating system inside an IBM z/VM® instance. The assignment of resources such as memory, processor cores, and HDDs to each virtual machine is done by the z/VM administrator. Because the BigInsights software is licensed “per node”, it is advantageous on the mainframe from a cost standpoint to have a relatively smaller number of nodes where each node has many cores, and many disk and memory resources that are assigned.

- Data replication: Because the HDFS is designed for large clusters of commodity systems, a key design assumption behind Hadoop is that servers and disk drives fail. However, a Linux cluster on the System z platform is highly stable, so data replication requirements might differ.

About IBM InfoSphere BigInsights

InfoSphere BigInsights is the IBM enterprise-grade Hadoop offering. It is based on industry-standard Apache Hadoop, but IBM provides extensive capabilities, including installation and management facilities and additional tools and utilities.

Special care is taken to ensure that InfoSphere BigInsights is 100% compatible with open source Hadoop. The IBM capabilities provide the Hadoop developer or administrator with additional choices and flexibility without locking them into proprietary technology.

Here are some of the standard open source utilities in InfoSphere BigInsights:

- PIG
- Hive / HCatalog
- Oozie
- HBASE
- Zookeeper
- Flume
- Avro
- Chukwa

Key software capabilities

IBM InfoSphere BigInsights provides advanced software capabilities that are not found in competing Hadoop distributions. Here are some of these capabilities:

- Big SQL: Big SQL is a rich, ANSI-compliant SQL implementation. Big SQL builds on 30 years of IBM experience in SQL and database engineering. Big SQL has several ad-vantages over competing SQL on Hadoop implementations:

  - SQL language compatibility
  - Support for native data sources
  - Performance
  - Federation
  - Security

  Unlike competing SQL-on-Hadoop implementations that introduce proprietary metadata or require that their own specific databases be deployed, Big SQL is open. Big SQL runs natively on existing Hadoop data sets in HDFS and uses existing Hadoop standards, such as the Hive metastore. For mainframe users, federation can be an important capability as well. From a single SQL query, users can query and combine data from multiple sources both within Hadoop clusters on or off the mainframe, including DB2 on z/OS databases with appropriate permissions.

- Big R: Big R is a set of libraries that provide end-to-end integration with the popular R programming language that is included in InfoSphere BigInsights. Big R provides a familiar environment for developers and data scientists proficient with the R language.

  Setting up a Big R environment requires some additional configuration in BigInsights for Linux on System z environments. Big R provides analytic functions that mirror existing language facilities. Big R implements parallelized execution across the Hadoop cluster, avoiding the need for developers to code their own MapReduce logic. A significant feature of Big R is that it supports downloadable packages from the Comprehensive R Archive Network. This allows these R packages to be “pushed down” and computed in parallel on the InfoSphere BigInsights cluster, which boosts performance and reduces effort on the part of developers.

  For more information about the R programming language, go to http://www.project-r.org.

- Big Sheets: Big Sheets is a spreadsheet style data manipulation and visualization tool that allows business users to access and analyze data in Hadoop without the need to be knowledgeable in Hadoop scripting languages or MapReduce programming. The BigSheets interface is shown in Figure 3. Using built-in line readers, BigSheets can import data in multiple formats. In this example, it is importing data that is stored in Hive.

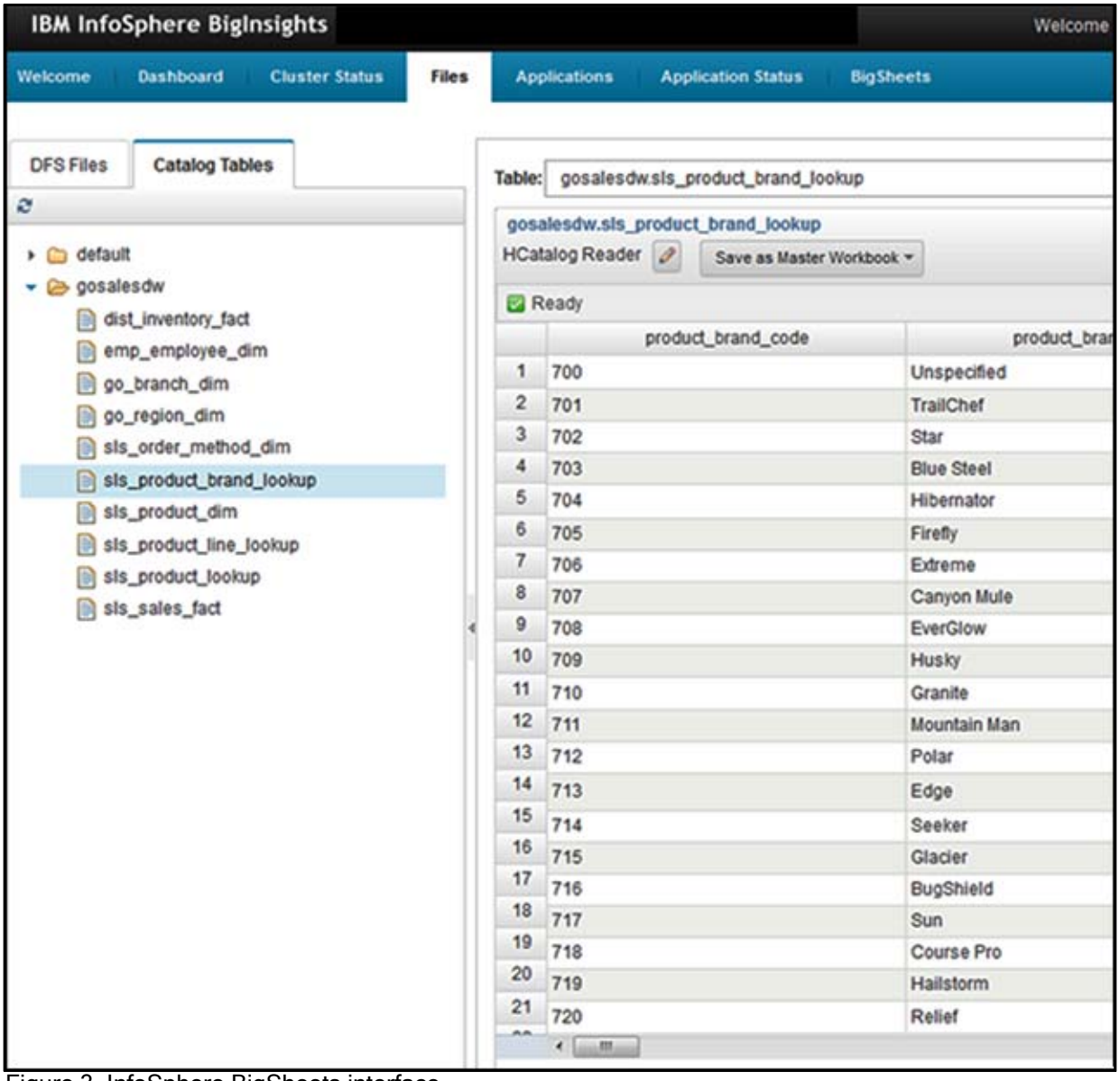

Figure 3. InfoSphere BigSheets interface

- Application Accelerators: IBM InfoSphere BigInsights extends the capabilities of open source Hadoop with accelerators that use pre-written capabilities for common big data use cases to build quickly high-quality applications. Here are some of the accelerators that are included in InfoSphere BigInsights:
    - Text Analytics Accelerators: A set of facilities for developing applications that analyze text across multiple spoken languages
    - Machine Data Accelerators: Tools that are aimed at developers that make it easy to develop applications that process log files, including web logs, mail logs, and various specialized file formats
    - Social Data Accelerators: Tools to easily import and analyze social data at scale from multiple online sources, including tweets, boards, and blogs

Not all capabilities that are available in InfoSphere BigInsights Enterprise Edition are supported on Linux for System z platforms. The following optionally deployable components are *not* supported by BigInsights for Linux on System z at this time:

- Adaptive MapReduce: An alternative, Hadoop-compatible scheduling framework that provides enhanced performance for latency sensitive Hadoop MapReduce jobs.
- IBM GPFS™ FPO: A variant of IBM GPFS that is POSIX-compliant and provides HDFS compatibility while providing Hadoop-style data locality by emulating the operation of the Hadoop NameNode.

Also, depending on the version of BigInsights that is deployed, some open source components might not be available on System z as a part of the BigInsights distribution.

Considerations when deploying on the mainframe

System z mainframes can interact with Hadoop clusters running on or off the mainframe. You might want to use Hadoop-based tools to analyze production data or combine other sensitive data with data from external sources and still operate within the security perimeter of the mainframe.

Hadoop, and BigInsights by extension, is a set of distributed software services that runs across multiple Linux compute hosts. BigInsights is well-suited to run within the IBM Linux for System z environment.

Although Linux for System z can run natively on mainframe central processors (CPs), in practice customers typically deploy Linux environments using the IBM Integrated Facility for Linux (IFL). By using the z/VM virtualization technology, one or more IFL processors can be allocated to logical partitions (LPARs). Each z/VM LPAR can have one or more Hadoop nodes where underlying system resources are mapped to the LPAR by using the IBM Processor Resource System Manager (IBM PR/SM™). As an alternative, resources from multiple LPARs can be aggregated and used by a single Hadoop cluster node. Storage resources can be similarly virtualized with storage coming from a mainframe-attached storage unit, such as an IBM System Storage® DS8000® series array. Because these subsystems provide all-flash storage or hybrid flash storage capabilities, they can typically deliver better I/O performance than local disk approaches that are normally used on commodity Intel clusters.

As shown in Figure 4, BigInsights clusters that are deployed on the mainframe typically are deployed across several nodes. Performance in Hadoop is a function of the system resources that are assigned to each node, and the number of nodes participating in the cluster. In traditional Hadoop clusters on Intel environments, total storage requirements also drive the size of the cluster. Total storage is less of a consideration in System z environments because the underlying storage is already virtualized.

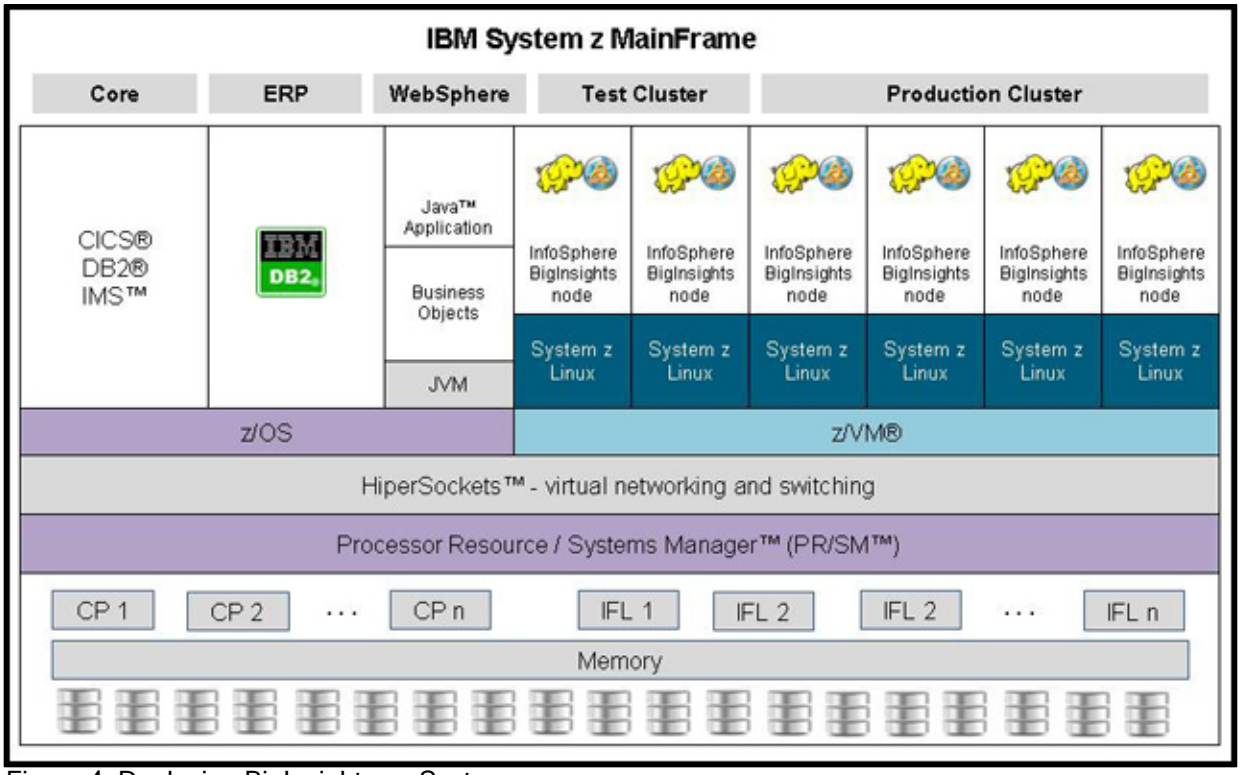

Figure 4. Deploying BigInsights on System z

Another mainframe advantage is that the underlying storage in a DS8000 series subsystem can be configured for reliability by using various RAID approaches to make sure that operations can continue in the event of the loss of a disk. Typically, Hadoop clusters rely on block replication to guard against the inevitable failure of disks. A block-replication factor of three is typical on Hadoop clusters, meaning that every block is replicated on three separate disks across the cluster. By avoiding or at least reducing the need for block replication because of the robustness and reliability of the platform, BigInsights on System z can make more efficient use of storage.

When deploying Hadoop clusters, normally at least one cluster host is allocated to the function of a master node, and other hosts are referred to as data nodes. In deploying a five-node cluster, for example, one node may be allocated to running key services, such as the BigInsights web console, HDFS NameNode Service, or JobTracker services. The data nodes in the cluster typically support the distributed HDFS file system and the various parallel frameworks on Hadoop, such has MapReduce, HBASE, and Big SQL.

As clusters become larger and cluster workload increases, it might be appropriate to deploy multiple master nodes. Multiple master nodes provide an additional degree of resiliency by allowing key Hadoop services to fail over to a second master node in the event of a hardware of software failure. It also provides more capacity because application managers (such as the Job Tracker) and workflow-oriented services (such as Oozie in Hadoop) normally run on master nodes instead of data nodes.

Additional deployment models

"Considerations when deploying on the mainframe" describes deployment of InfoSphere BigInsights directly on the System z. InfoSphere BigInsights also can be deployed on external clusters that are composed of Intel based systems or IBM Power Systems.

A third deployment model involves deploying InfoSphere BigInsights on a cloud-based service, such as IBM SoftLayer. This approach might be preferable when the following conditions apply:

- Clients do not want to be bothered with acquiring and maintain a Hadoop on cluster on their own premises.
- Data that is stored and processed on Hadoop comes largely from external sources as opposed to local data sources (such as cloud-based social media aggregators).
- Clients want to retain the flexibility to alter the size of clusters up and down based on changing requirements.

In the case of a System z hybrid Hadoop deployment that includes nodes running BigInsights on the IBM SoftLayer service, IBM provides a service that is called IBM BigInsights on Cloud. IBM typically provides a Vyatta appliance at the customer site along with the hosted Hadoop cloud service.

The Vyatta appliance is configured as a secure VPN tunnel between the client's network and the SoftLayer network supporting the Hadoop cluster. SoftLayer offers a "bare metal" infrastructure that can be used to build Hadoop clusters within SoftLayer. Bare metal servers support dedicated cloud-based clusters and provide the following advantages:

- Better quality of service because multiple virtualized environments are not sharing a physical infrastructure
- More predictable service levels because traditional problems such as "noisy neighbor" effects, which are problematic in competing cloud solutions, do not apply
- Better security isolation because clients customers are ensured access to their own dedicated compute, network, and storage infrastructure, which reduces the risk of data being compromised
- Client control of the entire compute environment

Additional information about deploying Hadoop and System z is available in *Hadoop and System z*, REDP-5142.

Usage scenarios

Customer that are running System z environments can augment the mainframe by using Hadoop in many different ways. This section looks at some of the more common usage scenarios for using InfoSphere BigInsights to help process mainframe data.

ETL processing

ETL processing is a common use case for Hadoop. Although offerings such as IBM InfoSphere DataStage® are well-suited for ETL requirements involving structured data on the mainframe, Hadoop might offer better ETL for other data types. Often, ETL is described as "ELT" in Hadoop environments, reflecting the fact that transformation operations are performed after data is loaded into Hadoop.

Performing ETL in Hadoop becomes more viable as data sets become large or when they can benefit from the parallel processing facilities that are inherent in Hadoop to reduce total processing times. Hadoop is well-suited to process semi-structured or unstructured data because Hadoop provides specific facilities for dealing with these types of data.

Analytic models that are composed of multiple data sources

For some types of business problems, clients might find themselves wanting to analyze data coming from multiple sources. As an example, data sources on z/OS might contain key information about customers and related transactional information.

To provide a complete view of the relationship between customers, it might be useful to combine customer transactional information with data coming from other sources:

- Logs from electronic self-service / self-provisioning systems
- Customer service interactions
- Web logs
- Text in emails, chats, or recorded telephone conversations with a call center
- Social media content that is gathered from various aggregators

To analyze all these data sources together, it is easier and more cost-efficient to move the data into a common “sandbox” where relevant data can be extracted and combined. Hadoop provides an ideal platform for this type of ad hoc data exploration. As relationships between data sources are better understood and some data views are deemed to be valuable, more robust processes can be devised to incorporate these additional data sources into analytic applications.

Ad hoc analysis of mainframe data

Experimenting directly on mainframe operational data is a bad idea. You risk corrupting data because of human or programming error, or causing adverse effects on the quality of service for production applications on z/OS.

In these cases, it might be preferable to take extracts of mainframe data from z/OS data sources and make the data available in native Hadoop formats, such as delimited data, Hive, or HBASE. Tools such as the IBM InfoSphere System z Connector for Hadoop provide an easy way to transfer efficiently large quantities of data from various mainframe data sources. Customers also can use other third-party solutions, such as the Dovetail Technologies Co:Z Co-Processing Toolkit.

By moving z/OS data to Hadoop for ad hoc analysis, you can experiment and create different views of data without adversely impacting operations. This is true on or off the mainframe. If particular views of data or new capabilities are deemed to be useful, they can be rolled into production applications on the mainframe, or the Hadoop environment can be used to produce these data views on an ongoing basis.

Mainframe log file analysis

Another important usage scenario for Hadoop and the mainframe is the analysis of various mainframe log file formats. IBM System Management Facility (SMF) is a component of z/OS for mainframe computers that provides a standardized method for writing out activity to a data set. SMF provides complete instrumentation of baseline activities on the mainframe, including I/O, network activity, software usage, processor usage, and other items. Add-on components, including DB2, IBM CICS®, IBM WebSphere® MQ, and IBM WebSphere Application Server, provide their own log file type reporting using SMF.

Hadoop, and BigInsights in particular, provide rich facilities for parsing, analyzing, and reporting on log file of all types. By analyzing logs using tools in BigInsights, clients can realize a number of benefits:

- Understand usage patterns by user, application, and group
- Identify issues before they affect production applications
- Gather trending information that is useful for capacity planning
- Find intrusion attempts, other security issues, or evidence of fraudulent activity

Because Hadoop is designed to support large data sets, clients can retain raw log data for longer periods than otherwise is feasible. More data helps clients discover long-term trends that are related to usage and variance in activities.

Extending DB2 functions by offloading processing to Hadoop

Firms are increasingly interested in incorporating external data into data models that are stored in production environments. A nice feature of DB2 for z/OS is that it provides specific facilities that allow user-defined functions to submit queries to Hadoop clusters on and off the mainframe, and read back the results of those queries and populate designated tables in DB2 for z/OS. This approach allows the mainframe to essentially “pull” summary information from trusted big data sources rather than needing to load and process all of the information in z/OS.

Integration

When extending the mainframe environment with Hadoop, a key consideration is how to integrate the environments and facilitate data movement. This section looks at some common ways to move data between the mainframe and Hadoop.

Triggering Hadoop analytics from IBM DB 2 for z/OS

DB2 for z/OS provides user-defined functions that simplify the integration between DB2 applications and InfoSphere BigInsights. Among these user-defined functions are JAQL_SUBMIT and HDFS_READ, which is a table-level function.

The basis of the integration is the use of JAQL, a query language that is primarily used for processing data in JavaScript Object Notation (JSON) format. JAQL can be used to perform efficiently operations on data in Hadoop, such as select, filter, join, and group. When processing large data sets, it is more efficient to pre-process data sets in Hadoop, and then move a smaller set of resulting data to the mainframe for processing. InfoSphere BigInsights provides a JAQL Server that allows remote entities (including DB2 for z/OS on the mainframe) to submit queries to Hadoop.

The JAQL_SUBMIT function enables users to start IBM InfoSphere BigInsights JAQL from a DB2 application. The related HDFS_READ function can read the contents of a file directly from HDFS and return the contents as a DB2 table.

HDFS_READ is useful either as a stand-alone function or when used with JAQL_SUBMIT. As a use case for the former, if a customer has a data source in Hadoop that is being periodically updated automatically, the customer can use HDFS_READ in DB2 applications to make sure that mainframe applications always reflect the latest available data from the Hadoop cluster. Using JAQL_SUBMIT, customers can run a JSON Query on Hadoop to generate the file before reading the results into DB2.

This capability is described in detail in the InfoSphere BigInsights documentation. The following sample select statement in DB2 shows where a table is generated dynamically by the query based on a JAQL script that is submitted to BigInsights, the results of which are returned to DB2 by the HDFS_READ function:

```
SELECT * FROM temp, TABLE(hdfs_read(jaql_submit(
  'expandFD([$id_TMP,$accounts_TMP,$balance_TMP,$desc_TMP]
    ->write(del("lsd.out",
      { schema: schema
        { id:long,account:decfloat,balance:double,desc:string
        }
      }));',temp.params,'9.125.91.40','8080','600000')))
  AS Write1000Row;
```

Using Big SQL to access to multiple data sources

Big SQL in InfoSphere BigInsights supports the federation of multiple data sources, including (but not limited to) DB2, IBM PureData® System for Analytics (Netezza), IBM PureData System for Operational Analytics, Teradata, and Oracle.

Whereas the approach that is described in "Triggering Hadoop analytics from IBM DB2 for z/OS" is about DB2 “pulling” information from BigInsights, this approach is different. It involves Big SQL being used to pull information from Hadoop, DB2, and other data sources before optionally storing a combined result on the BigInsights cluster.

To users and client applications, data sources appear as a single collective group in the Big SQL server. Users and applications interface with Big SQL to access the data. The Big SQL server contains a system catalog that stores information about the data that it can access. This catalog contains entries that identify data sources and their characteristics.

IBM InfoSphere System z Connector for Hadoop

Another important tool for integrating InfoSphere BigInsights with System z data sources is the IBM InfoSphere System z Connector for Hadoop.

IBM InfoSphere System z Connector for Hadoop provides fast and seamless data connectivity between various mainframe data sources and IBM InfoSphere BigInsights. You can easily extract data from z/OS sources, including IBM DB2 for z/OS, IBM IMS, VSAM, and various system and application log files, without the need for mainframe-based SQL queries, custom programming, or specialized skills. When data is in Hadoop, you can use the rich capabilities of IBM InfoSphere BigInsights to process and analyze data quickly and cost-efficiently. Hadoop processing can take place on an external cluster that is connected to the IBM zEnterprise® mainframe or directly on a mainframe running BigInsights Linux on System z.

By extending the capabilities of System z with IBM InfoSphere BigInsights and the IBM InfoSphere System z Connector for Hadoop, users can create a hybrid transaction and analytic processing platform that can manage mixed workloads, while maintaining the high quality of service, security, and integrity that IBM mainframe users have come to expect.

Running analytic tools to access Hadoop data

In addition to technical integration approaches, System z and other platforms may also run additional applications that directly support connectivity to Hadoop clusters. Examples of analytics offerings in the IBM portfolio that support processing on Hadoop include IBM SPSS®, IBM Cognos®, and IBM Tealeaf®.

Many third-party applications, including SAS and MicroStrategy, also support IBM InfoSphere BigInsights. These ISV applications might run either on or off the mainframe and, depending on the application, might be supported in Linux for System z partitions.
The specifics of integration approaches and supported versions vary by ISV.

Supported platforms

At the time of writing, InfoSphere BigInsights V2.1.2 is certified on Linux on System z. It is supported on Red Hat Enterprise Linux 6.4 running on a z/VM environment or on native Linux on System z LPARs. It is not supported on SLES environments.

BigInsights on the mainframe normally is deployed on IFL processors on System z environments.

For more information about how RHEL 6.4 is deployed on IBM z/VM environments, see *The Virtualization Cookbook for IBM z/VM 6.3, RHEL 6.4, and SLES 11 SP3*, SG24-8147.

Ordering information

IBM InfoSphere BigInsights on System z is licensed per node. A node refers to a node running on a virtual machine running an instance of an operating system.

With Hadoop clusters, it is typical to have dense nodes that are composed of many cores and multiple individual HDDs. Having fewer relatively dense compute and storage cluster nodes (RHEL OS instances) is more cost-efficient to operate because it reduces the number of InfoSphere BigInsights licenses that must be purchased. In Intel based Hadoop cluster environments, it is typical to have 12 - 16 cores per cluster node with a similar number of individual dedicated HDDs.

InfoSphere BigInsights Enterprise Edition is licensed per node or per virtual server (the latter being the appropriate metric in System z environments). The InfoSphere BigInsights Standard Edition is not available for ordering on System z.

Here are the descriptions and part numbers for orderable components:

- D13E9LL: InfoSphere BigInsights Enterprise Edition for System z License + SW sup-port and subscription services for 12 months
- E0J67LL: InfoSphere BigInsights Enterprise Edition for System z annual SW support and subscription services renewal for 12 months
- D13EALL: InfoSphere BigInsights Enterprise Edition for System z annual SW support and subscription services reinstatement

Related platforms

IBM InfoSphere BigInsights also can be deployed on other platforms, including IBM and competitive Intel based systems, IBM Power Systems, and cloud service offerings, including IBM SoftLayer and third-party infrastructure-as-a-service (IaaS) offerings.

Customers that want to simplify the process of extracting and translating data from mainframe sources into Hadoop formats might want to use the IBM InfoSphere System z Connector for Hadoop, which is available as a separate offering. The System z Connector for Hadoop automates the process of transferring data from mainframe z/OS sources, including VSAM, QSAM, IMS, DB2, and various log file formats, and stores data in various formats on BigInsights Hadoop clusters on System z or clusters off the mainframe.

Related information

- *Hadoop and System z*, REDP-5142
- *The Virtualization Cookbook for IBM z/VM 6.3, RHEL 6.4, and SLES 11 SP3*, SG24-8147
- IBM InfoSphere BigInsights
  ibm.com/software/os/systemz/biginsightsz
- R programming language
  http://www.project-r.org

Notices

This information was developed for products and services offered in the U.S.A.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service. IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing, to:

IBM Director of Licensing, IBM Corporation, North Castle Drive, Armonk, NY 10504-1785 U.S.A.

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law: INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you. This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this IBM product and use of those Web sites is at your own risk.IBM may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you. Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products. This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

Any performance data contained herein was determined in a controlled environment. Therefore, the results obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurement may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs.

This document was created or updated on November 26, 2014.

Send us your comments in one of the following ways:

- Use the online **Contact us** review form found at:
  ibm.com/redbooks
- Send your comments in an e-mail to:
  redbooks@us.ibm.com
- Mail your comments to:
  IBM Corporation, International Technical Support Organization
  Dept. HYTD Mail Station P099
  2455 South Road
  Poughkeepsie, NY 12601-5400 U.S.A.

This document is available online at http://www.ibm.com/redbooks/abstracts/tips1215.html .

Trademarks

The following terms are trademarks of other companies:

Netezza, and N logo are trademarks or registered trademarks of IBM International Group B.V., an IBM Company.

SoftLayer, and SoftLayer device are trademarks or registered trademarks of SoftLayer, Inc., an IBM Company.

Intel, Intel logo, Intel Inside logo, and Intel Centrino logo are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

Linux is a trademark of Linus Torvalds in the United States, other countries, or both.

Java, and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.

Other company, product, or service names may be trademarks or service marks of others.