#SHAREorg

SHARE
Technology · Connections · Results

Getting Started with Big Data Analytics for the Enterprise

Mike Biere
IBM

Tuesday August 7th
Session Number : 11930

IBM PRESS AUTHOR

SHARE in Anaheim 2012

IBM's Big Data Portfolio

IBM views Big Data at the enterprise level thus we aren't honing in on one aspect such as analysis of social media or federated data. Our solutions span 4 key areas:

1. Data Warehouse (Information Server, DB2 Analytics Accelerator, Neteeza, etc.)
2. BigInsights (Hadoop etc.)
3. Stream data capture and analysis
4. Federated data discovery and analysis

What can you do with big data?

Act on Deeper Customer Insight

- Social media customer sentiment analysis
- Promotion optimization
- Segmentation
- Customer profitability
- Click-stream analysis
- CDR processing
- Multi-channel interaction analysis
- Loyalty program analytics
- Churn prediction

Create Innovative New Products

- Social Media - Product/brand Sentiment analysis
- Brand strategy
- Market analysis
- RFID tracking & analysis
- Transaction analysis to create insight-based product/service offerings

Optimize your Operational Processes

- Smart Grid/meter management
- Distribution load forecasting
- Sales reporting
- Inventory & merchandising optimization
- Options trading
- ICU patient monitoring
- Disease surveillance
- Transportation network optimization
- Store performance
- Environmental analysis
- Experimental research

Prevent Fraud and Reduce Risk

- Multimodal surveillance
- Cyber security
- Fraud modeling & detection
- Risk modeling & management
- Regulatory reporting

Proactively Maintain your Assets

- Network analytics
- Asset management and predictive issue resolution
- Website analytics
- IT log analysis

Pains Addressed by a Big Data Platform

- High cost of storing and analyzing data combined with data growing volumes
- Cost and performance of enterprise data warehouse - single DW cannot meet everyone's needs
- Inability to exploit new sources of data – need to explore, prove value, and extract it cost effectively
- Loss of fidelity and huge time/cost to convert unstructured data (video, audio, textual content) to structured format for analysis
- Inability to act and high cost of acting on data in real-time leads to lost opportunities
- High cost to maintain data online when it could exist in an online archive – query-able archive

IBM Big Data Strategy: move the analytics closer to the data

New analytic applications drive the requirements for a big data platform

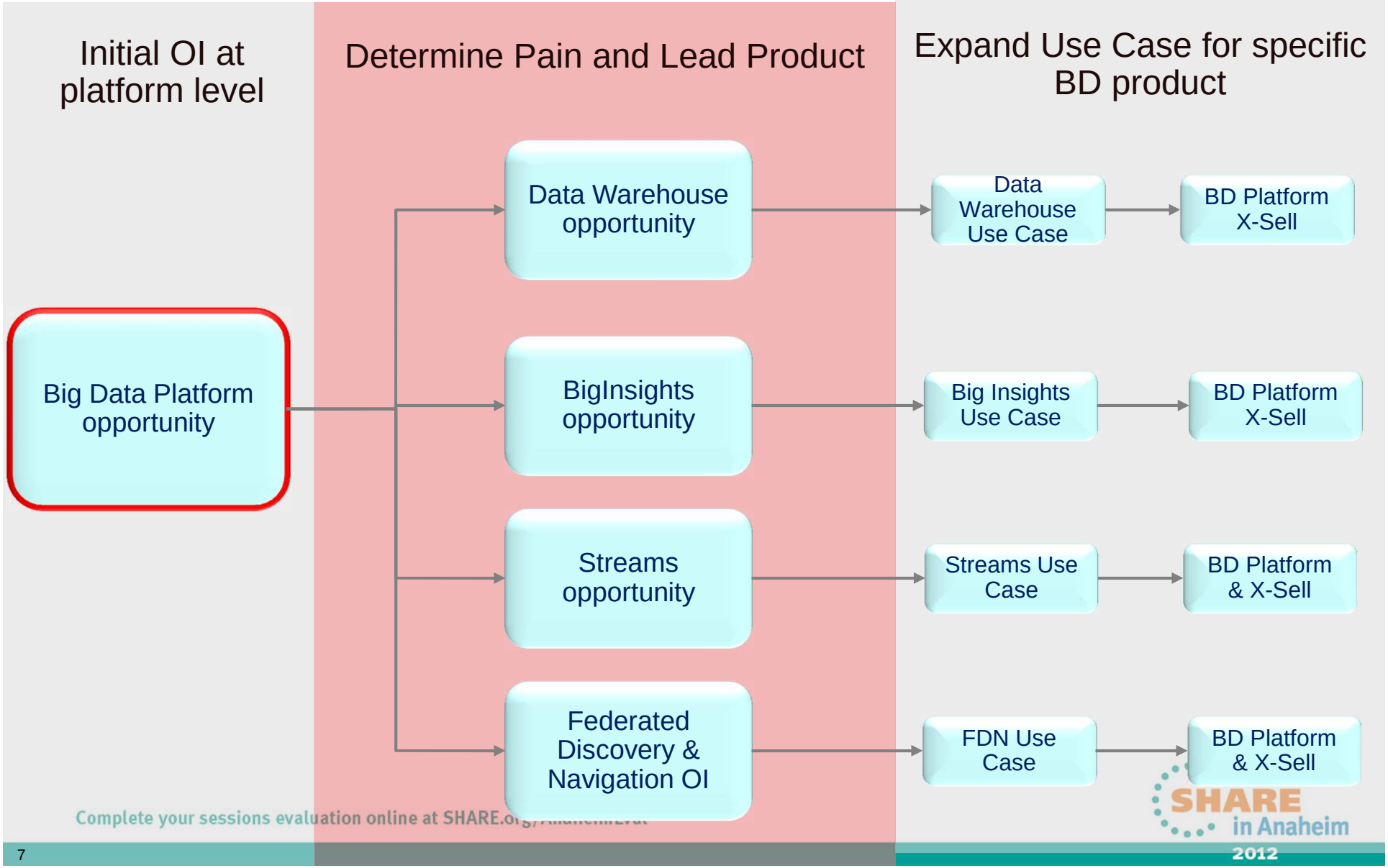
- Integrate and manage the full variety, velocity and volume of data
- Apply advanced analytics to information in its native form
- Visualize all available data for ad-hoc analysis
- Development environment for building new analytic applications
- Workload optimization and scheduling
- Security and Governance

Big Data Platform – a consultant's view

Diagnose → Prescribe

Platform	Determine Pain and Leading Product	Expand Use Case for BD product x-sell
1. Understand their data assets 2. Current use of their data assets and planned future use 3. Understand new sources and combinations of data that has business impact 4. Message Big Data Pain Points and Use Cases	1. Message IBM thought leadership on leveraging their existing data assets and the BD platform 2. Identify the granular pain point 3. Determine the business case for Big data 4. Determine Lead product 5. Further qualify that product	1. Determine the use case 2. Position the big data platform (complimentary products for the use case) 3. Determine cross-sell potential for BD products

Big Data Platform

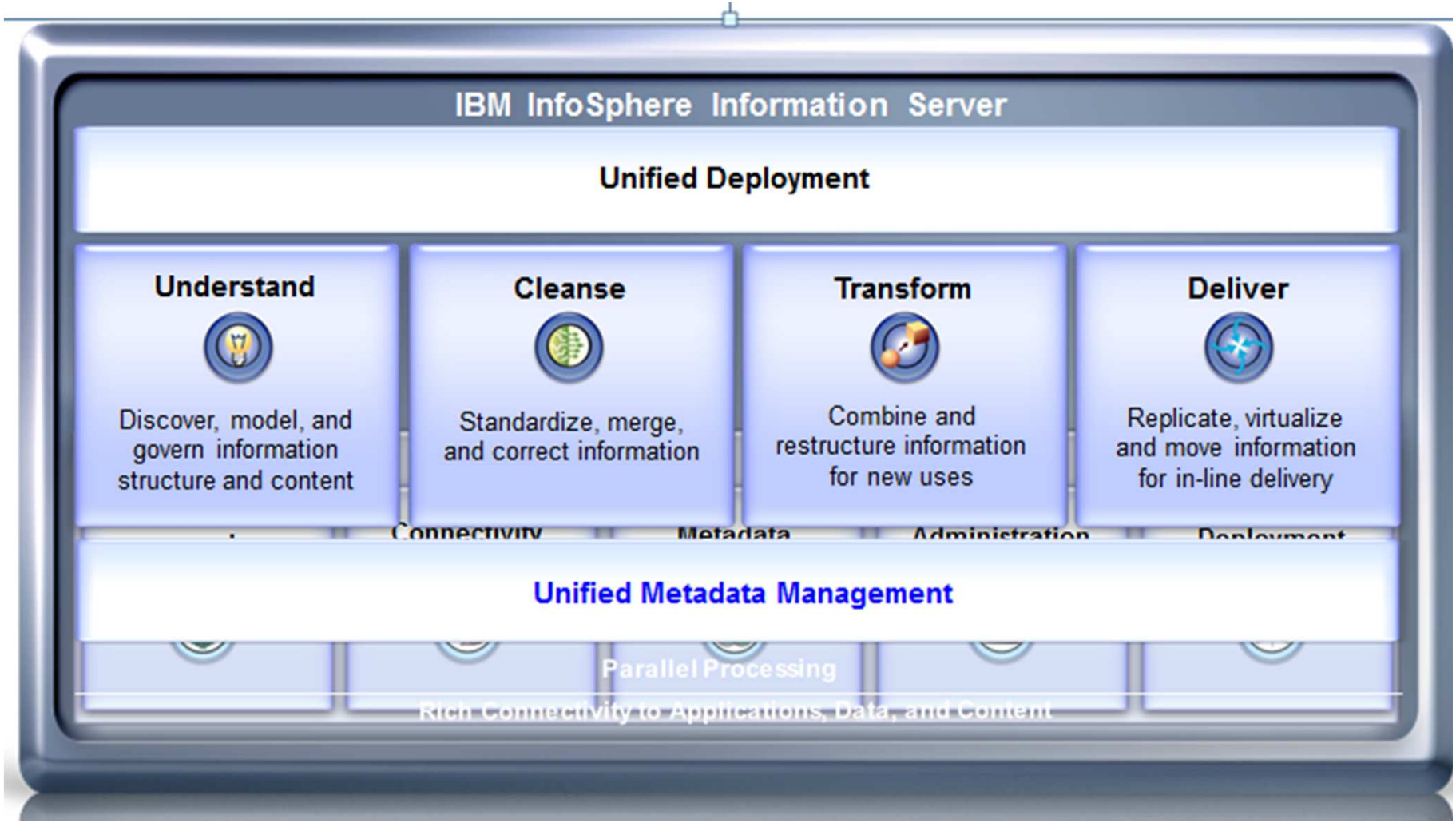


Big Data Platform – self assessment

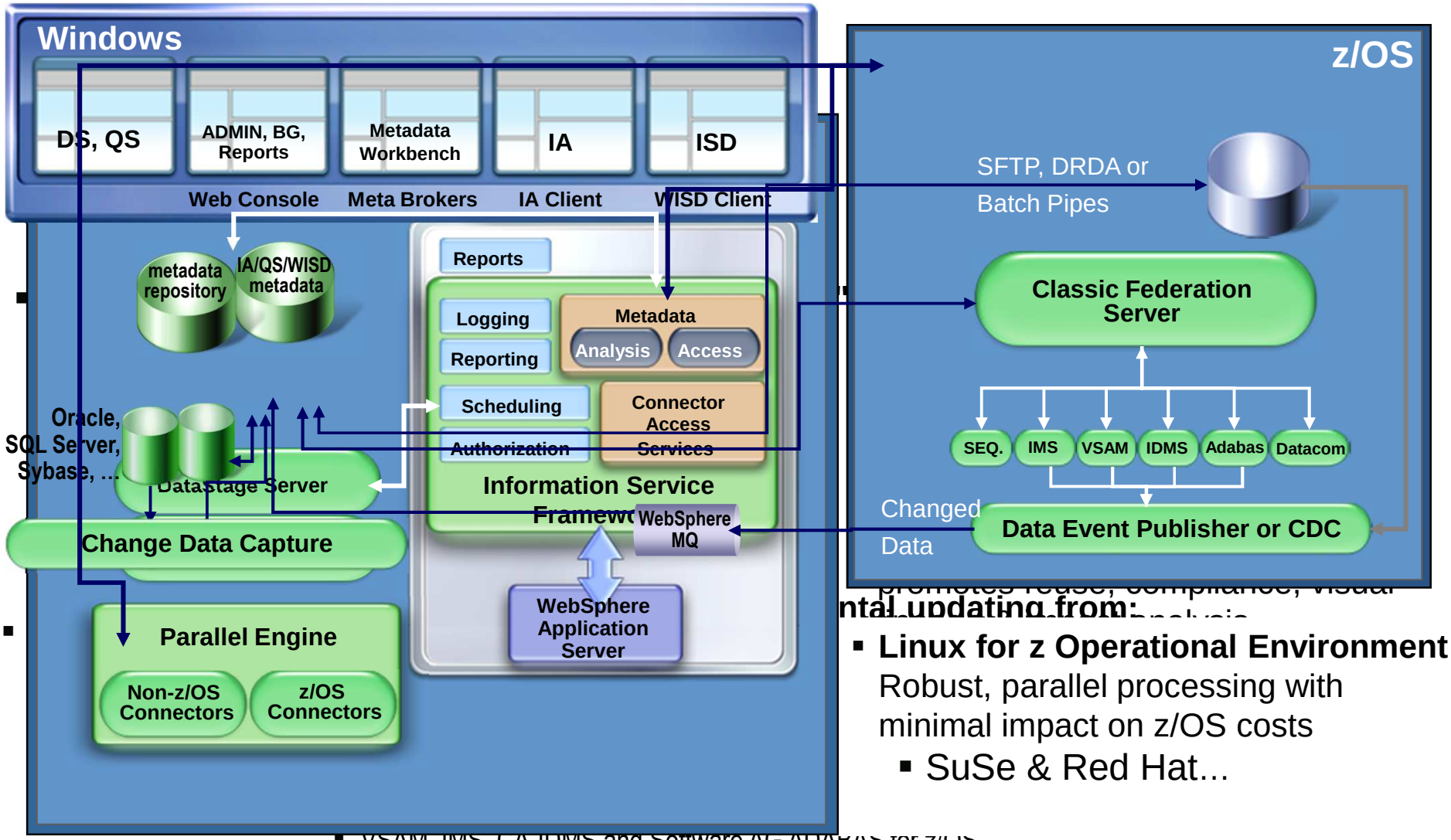
Questions	Pain	Lead Product
• Do you have performance challenges with your DW? High number of concurrent users/queries? • Do you expect your query/user volume to grow? • Is the volume in your DW increasing (TBs and PB)?	• Too much latency for user queries to DW • Volume of structured information is growing and straining performance	Data Warehousing
• Do you want to analyze both structured and unstructured data together, without converging them to one schema? • Are there any projects where you do not analyze the full volume of data available to you? Why not? • Are you concerned with the cost of managing growing data volumes in traditional technology?	• Inability to analyze a variety of data in its native format • Persisting and analyzing all available data results in poor performance or huge costs	InfoSphere BigInsights
• Do you have the need to analyze data in real-time? • Would you like to analyze a body of data that is simply too large to persist in any technology?	• Inability to analyze data in motion resulting in too much latency in insight • Too costly to store and analyze all available data	InfoSphere Streams

Big Data Platform – Data Warehouse

InfoSphere Information Server



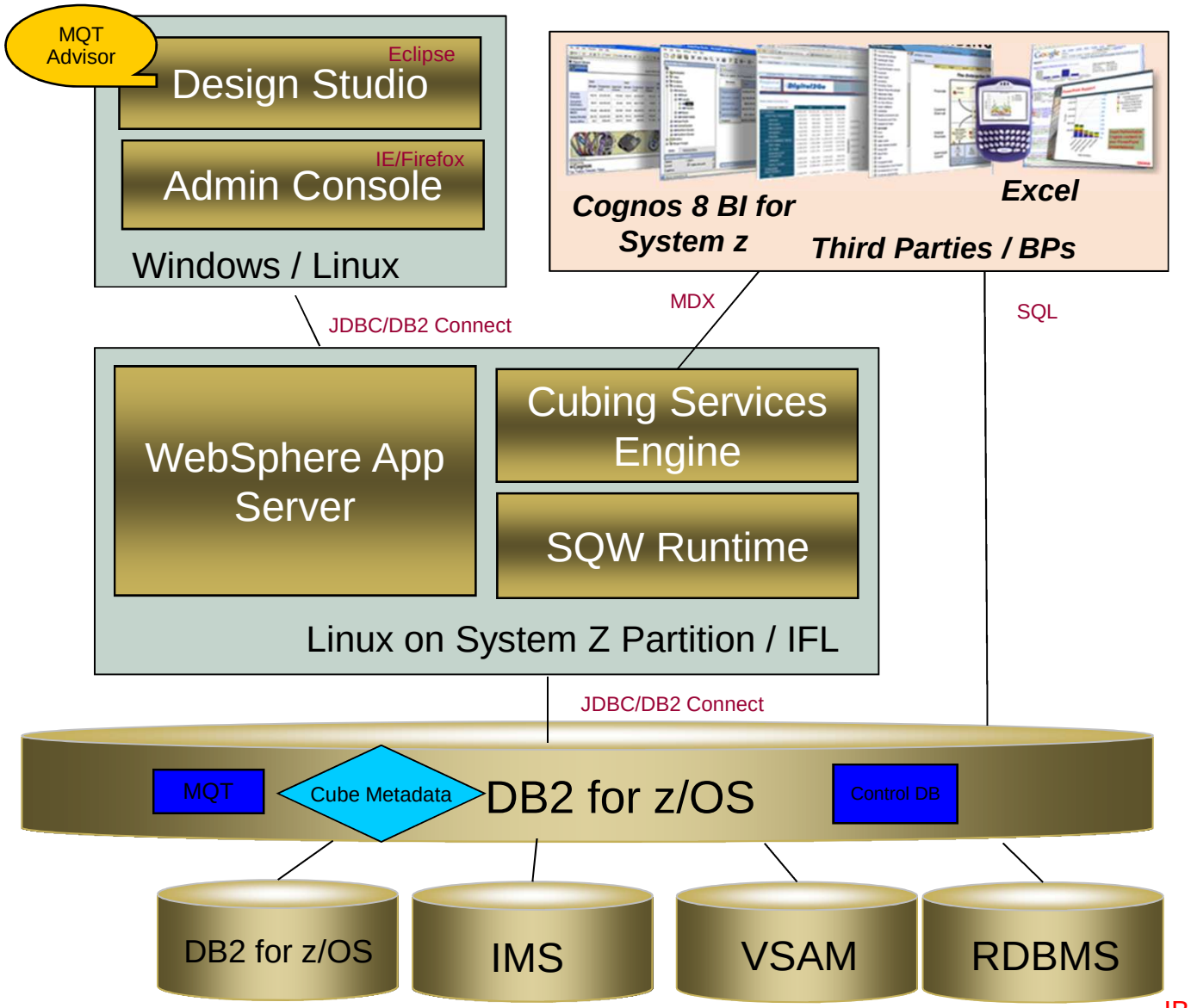
Information Server and Foundation Tools for System z



- **Linux for z Operational Environment**
  Robust, parallel processing with minimal impact on z/OS costs
  - SuSe & Red Hat...

- VSAM, IMS, CA-IDMS and Software AG ADABAS for z/OS

InfoSphere Warehouse for DB2 for z/OS



Client Layer

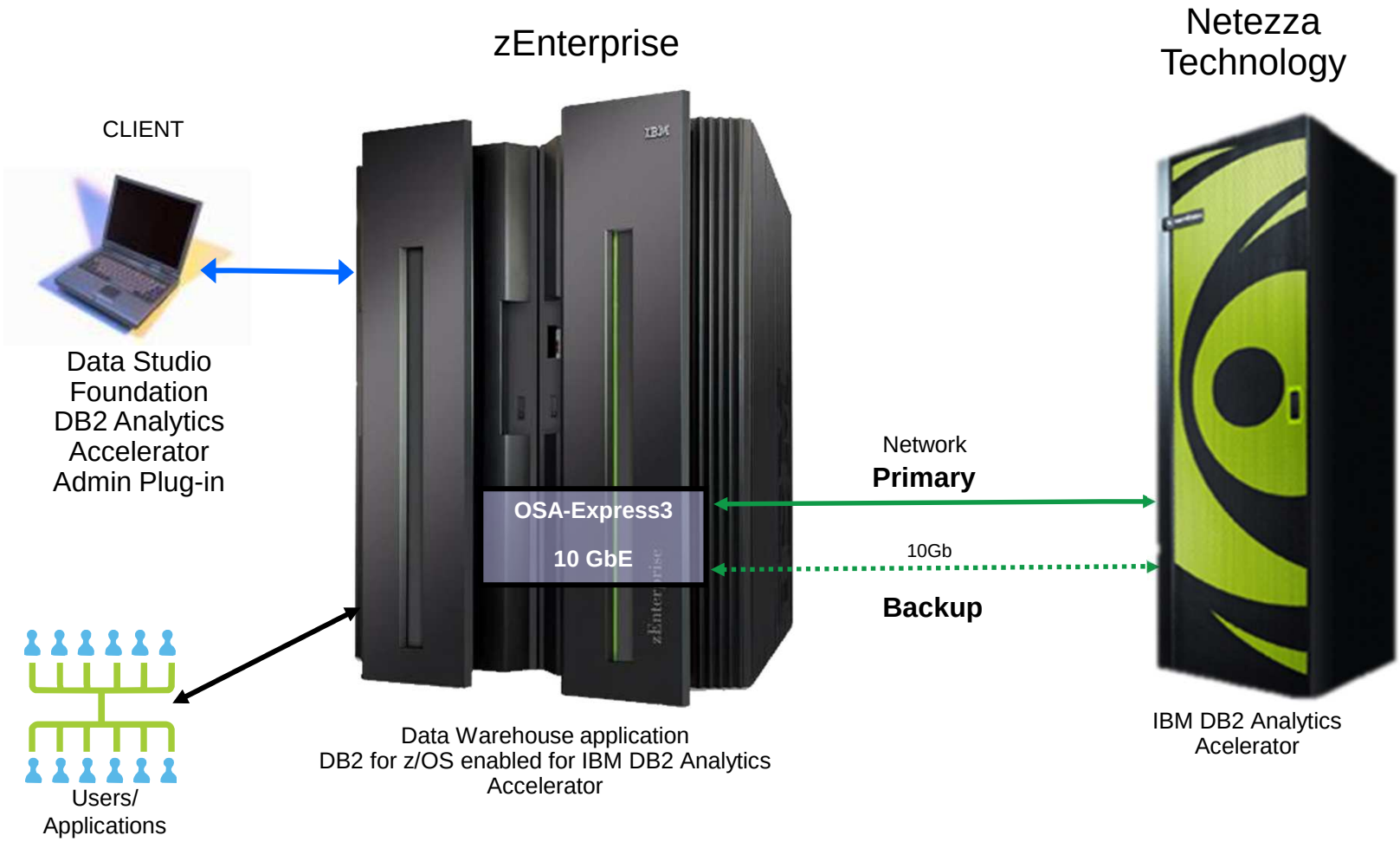
- Design and admin client
- BI / Reporting tools and apps

Application Server

Data Warehouse Server

Source Systems

IBM DB2 Analytics Accelerator V2 product components



Performance & Savings

Query	Total Rows Reviewed	Total Rows Returned	DB2 Only Hours	DB2 Only Sec(s)	DB2 with IDAA Hours	DB2 with IDAA Sec(s)	Times Faster
Query 1	2,813,571	853,320	2:39	9,540	0.0	5	1,908
Query 2	2,813,571	585,780	2:16	8,220	0.0	5	1,644
Query 3	8,260,214	274	1:16	4,560	0.0	6	760
Query 4	2,813,571	601,197	1:08	4,080	0.0	5	816
Query 5	3,422,765	508	0:57	4,080	0.0	70	58
Query 6	4,290,648	165	0:53	3,180	0.0	6	530
Query 7	361,521	58,236	0:51	3,120	0.0	4	780
Query 8	3,425.29	724	0:44	2,640	0.0	2	1,320
Query 9	4,130,107	137	0:42	2,520	0.1	193	13

Queries run faster

- Save CPU resources
- People time
- Business opportunities

Actual customer results, October 2011

Aetna™

DB2 Analytics Accelerator: "we had this up and running in days with queries that ran over 1000 times faster"

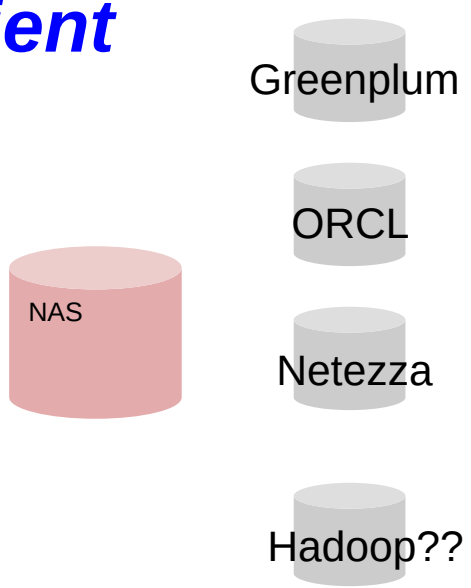
Swiss Re

DB2 Analytics Accelerator: "we expect ROI in less than 4 months"

Accelerating decisions to the speed of business

Big Data Platform- BigInsights

Financial Industry Regulatory client



- Backdrop
  - 20b records/day processed… 6tb/day
  - Data growing 80%/year
  - Client refuses to see their costs grow 80% year
  - Goal: introduce a Hadoop store, to reduce costs

- Plan
  - The short term – examine BigInsights into their environment, while allowing them to easily leverage current frontends (Cognos, apps, custom web interfaces, etc)
  - They are interested in 3 accelerators: text, stats, predictive
  - Long term: they will move data segments out of traditional warehouse/databases and into Hadoop

Profile of a BigInsights solution

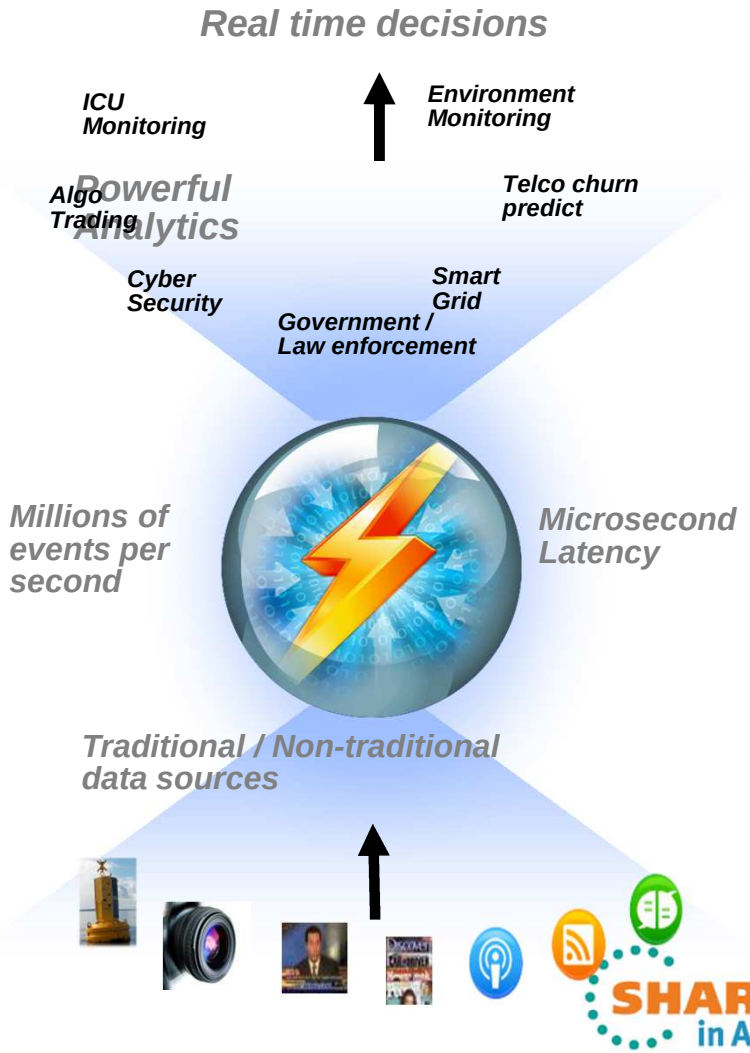
- Very large data sets (TBs to PBs)
- Schema-less data in native format
- Low user concurrency
- Open source, non-proprietary solution
- Support for non SQL development tools (MapReduce, R)
- Need to explore data with questions you can't anticipate
- Analytics across and match unstructured and non-standard data types
- Store data once but "look at in multiple ways" – i.e. multiple data structures
- Desire to analyze data in place, without moving it or loading it
- Analytical sandbox to explore data, outside the organization's "official" restricted-access data management platforms
- Large data archive that you want available for occasional query and reporting access, but which is not valuable enough to host in a warehouse

Big Data Platform- Streams

IBM InfoSphere Streams v2.0

A platform for real-time analytics on BIG data

- Volume
  - Terabytes per second
  - Petabytes per day
- Variety
  - All kinds of data
  - All kinds of analytics
- Velocity
  - Insights in microseconds
- Agility
  - Dynamically responsive
  - Rapid application development



Why InfoSphere Streams?

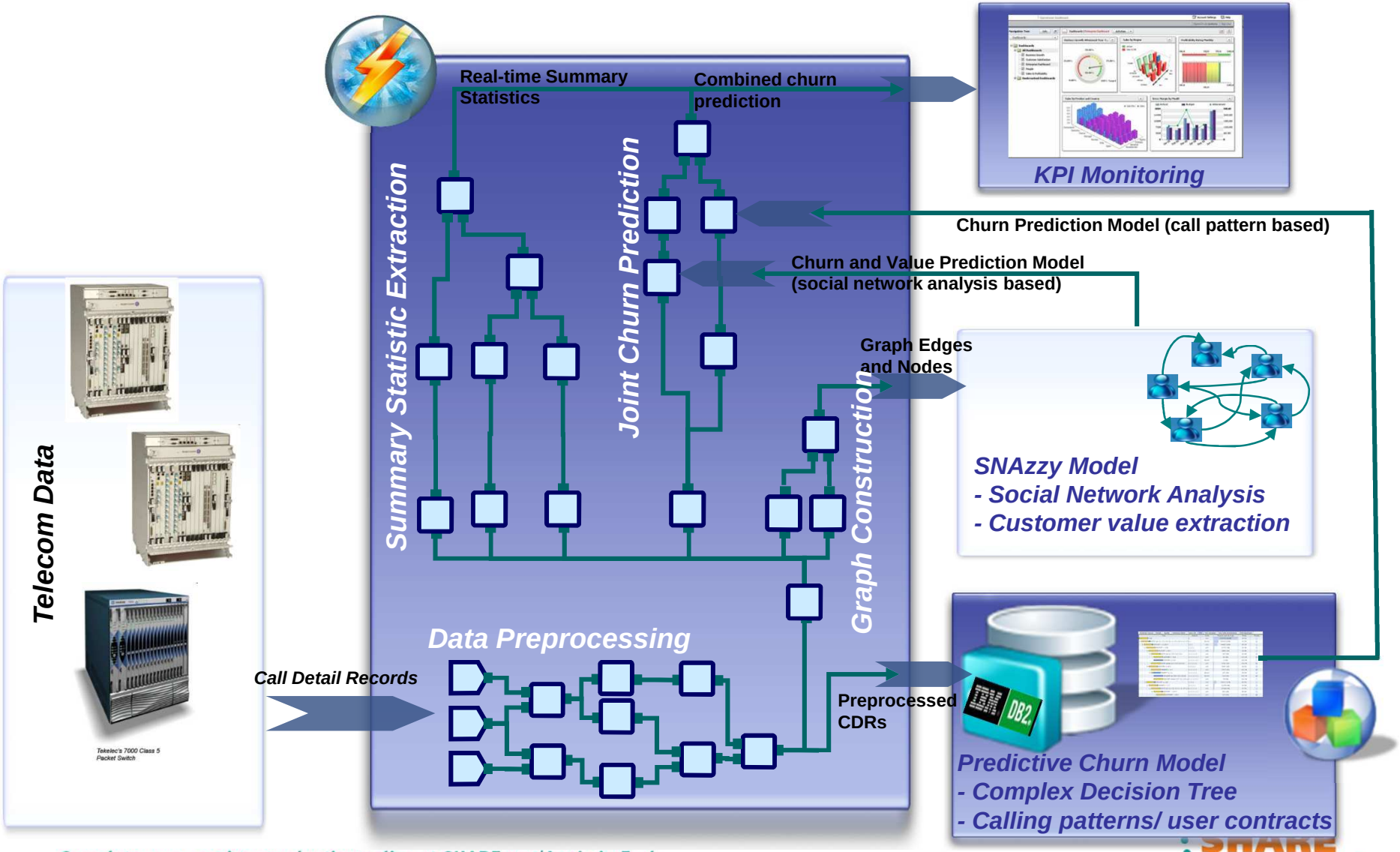
- Applications that require on-the-fly processing, filtering and analysis of streaming data
  - Sensors: environmental, industrial, surveillance video, GPS, …
  - "Data exhaust": network/system/web server/app server log files
  - High-rate transaction data: financial transactions, call detail records

- Criteria: two or more of the following
  - Messages are processed in isolation or in limited data windows
  - Sources include non-traditional data (spatial, imagery, text, …)
  - Sources vary in connection methods, data rates, and processing requirements, presenting integration challenges
  - Data rates/volumes require the resources of multiple processing nodes
  - Analysis and response are needed with sub-millisecond latency
  - Data rates and volumes are too great for store-and-mine approaches

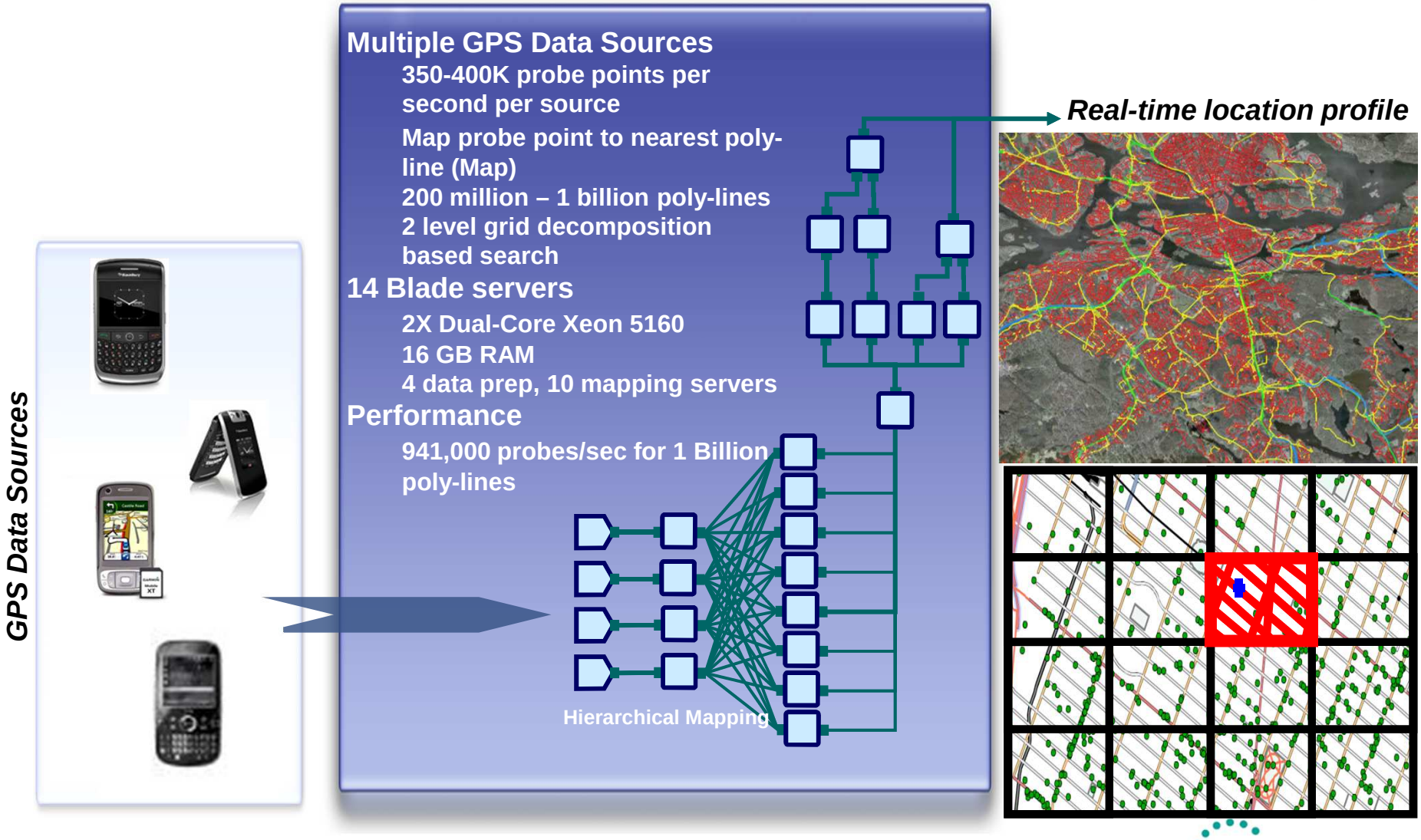
Streams usage case

- Data in motion, streaming data
- The value ‘shelf life’ of the data is narrow
- Volume + Speed + Analytic Requirement = Performance Challenge
- Structured, unstructured or non conventional data types
- Bring analytics to data, not data to the analytics
- Real time scoring using predictive models or rules based engine
- Need to examine and respond to information in real time
- Can’t ingest, examine and respond to the new high speed, high volume data sources hitting my existing DM and DW solutions
- React sooner to reduce risk, detect and prevent fraud, or prevent dangers (national security, power plant)
- Notice events sooner to capture sales opportunities, connect with in market customers, or notice patterns that matter to my business.
- Respond to new information in flight, in real time - before it lands

Telephony Architecture



Streams for Real-Time Geomapping



Big Data Platform- Federated discovery

Identifying a Vivisimo opportunity – what a consultant looks for

Top Use Cases

1. Understanding and viewing/navigating big data sources before importing data to a Hadoop system
   - Navigate, discover, and view big data sources to understand their potential value
2. Searching and navigating federated big data sources for operational applications
   - Customer service
   - Sales / Product recommendation
   - Order management

Qualifying Questions

1. How many sources of big data do you have in your business?
2. Do you understand the potential value of those big data sources today?
3. Do you have a need to discover / preview those data sources before importing/analyzing them?
4. Are you wondering how to get started with big data and you're unsure which data to import into a Hadoop system?
5. Do you have operational processes that require people to search multiple repositories of varied content? How much time is wasted with those manual processes?

Vivisimo overview – value proposition & differentiators

Value Proposition

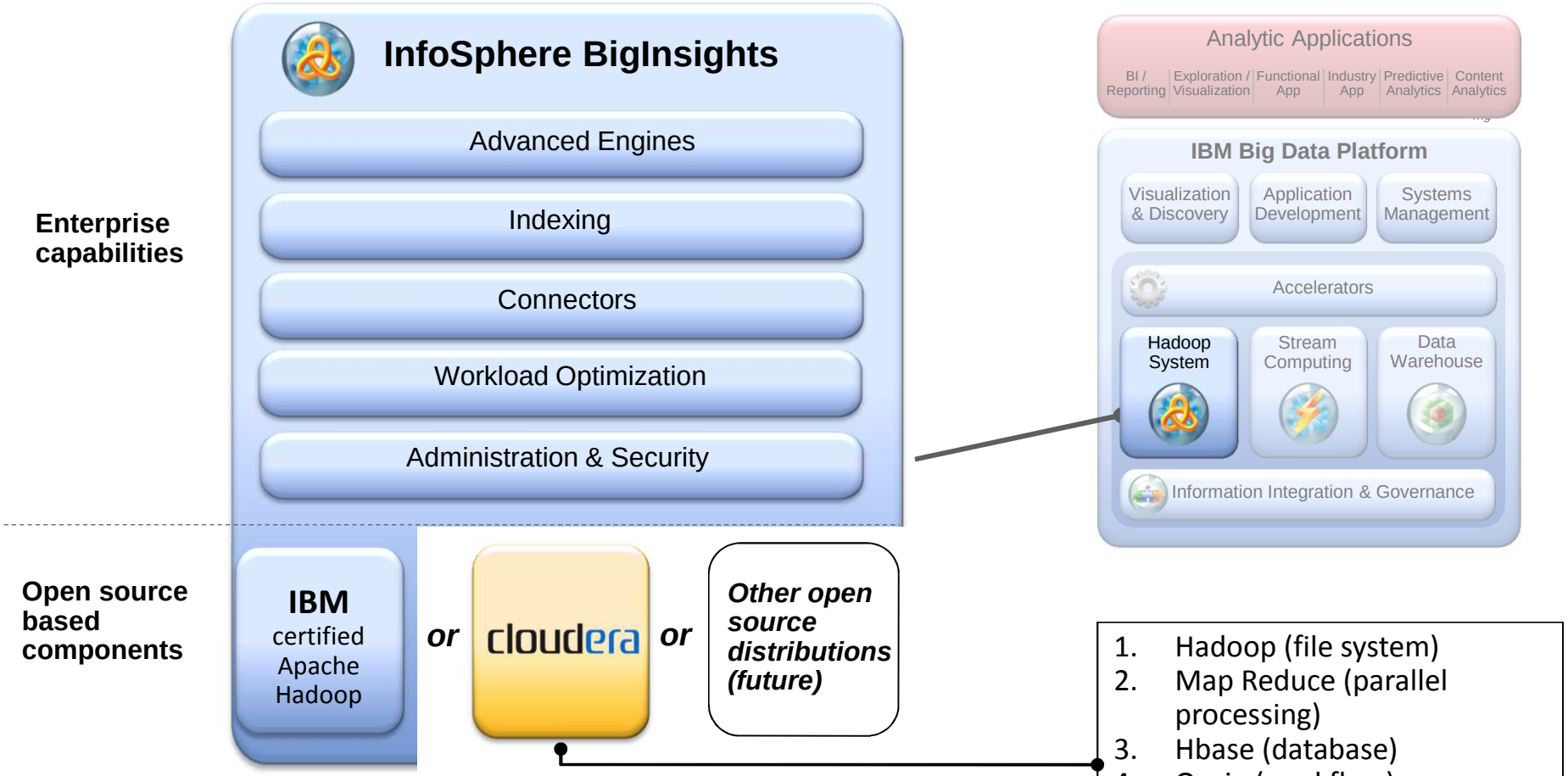
Accuracy – more relevant results due to position-based indexing
Security – respects the security rights of underlying systems
Scalability – scales to trillions of records

Differentiators

- **Unique Federated Discovery and Navigation Technology**
  - Position-based vs. vector-based index
  - Clustering and faceting to navigate data results
- **Scalable Architecture**
  - Fully distributed, fault-tolerant , unlimited scalability
- **Advanced On-the-Fly Analytics**
  - State-of-the-art real-time text and meta-data analytics
- **Secure Connectivity**
  - Secure data integration of multiple repositories in complex IT environments
- **Powerful Development Tools**
  - Easy-to-deploy applications across varied and large data sets & sources
- **Fast Time to Value**
  - Rapid deployments from POCs to production

Vivisimo Product Overview

IBM's Big Data platform will support open source distributions supporting multiple client bases



Use Case at a Current IBM Opportunity

Improve customer satisfaction and lower costs

- Problem
  - Gain 360 customer view of customer to offer optimal and relevant services customized to the customer' needs
  - Information locked into multiple data sources (ERP, Teradata, Mainframes, Social Media Content, Call Center Apps, Homegrown Apps, etc.) in the form of both structured and unstructured data)
- Solution
  - faceted exploration to explore all repositories, extract and index all metadata
  - Extract relevancy and support insurance industry - specific ontology for text analytics
  - Provide connectivity to complex internal repositories such as ERP, mainframe systems, as well as external data, such as web, email, feeds, web 2.0 social media content
- Usage and applicability of Vivisimo
  - Vivisimo can provide single point access to all data sources without copying the data into Hadoop
  - Ability to combine disparate data sources increases employee productivity and lowers costs
  - BigInsights can be used to run deep analytic jobs and Vivisimo can provide adhoc answers
  - Multiple front-end applications leveraging common back-end infrastructure

High availability and manageability for long analytic computing threads

Investigative analytic applications leverage Vivisimo velocity platform suited for navigation

BigInsights for Batch Processing Hadoop

Vivisimo for Adhoc analysis

Vivisimo Connectors crawl, index and load

Unstructured Data (email, Calls center logs, etc)

Content Mgmt Systems

Teradata Warehouse

Mainframes

ERP, CRM

- **Customer profiles** (demographics, relevancy amongst family members and friends, product sentiments, needs)
- **Customer call details**
- **Client email exchanges**
- Lessons learned

Summary

- Where are you in the range of Big data solutions requirements?
  - ✓ Data Warehouse not in order? … start here
  - ✓ Big Data team/project office/competency center in place? No … better get a plan in place
  - ✓ Data for analysis at all levels identified? No … an absolute necessity!!
  - ✓ Enterprise level approach or fit for purpose approach? FFP … better discuss how these may need to dovetail later and you'll be starting over.
  - ✓ Know what others in your industry are doing? No? … those who hesitate can be outperformed 2-3x
  - ✓ Identify and remember the implications of social media upon your business. It isn't about 'kids' tweeting anymore but you need to separate the 'noise' from the real information.