Big Data und Hadoop

BI/DW Modernisierungs-Szenarien auf System z

Eberhard Hechler
Executive Architect, Member IBM Academy of Technology
IBM Germany R&D Lab
Trademarks

The following are trademarks of the International Business Machines Corporation in the United States and/or other countries.

* Registered trademarks of IBM Corporation

The following are trademarks or registered trademarks of other companies.

Adobe, the Adobe logo, PostScript, and the PostScript logo are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, and/or other countries.

IT Infrastructure Library is a registered trademark of the Central Computer and Telecommunications Agency which is now part of the Office of Government Commerce.

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

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

Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

Windows Server and the Windows logo are trademarks of the Microsoft group of countries.

ITIL is a registered trademark, and a registered community trademark of the Office of Government Commerce, and is registered in the U.S. Patent and Trademark Office.

UNIX is a registered trademark of The Open Group in the United States and other countries.

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

Cell Broadband Engine is a trademark of Sony Computer Entertainment, Inc. in the United States, other countries, or both and is used under license therefrom.

Linear Tape-Open, LTO, the LTO Logo, Ultrium, and the Ultrium logo are trademarks of HP, IBM Corp. and Quantum in the U.S. and other countries.

* Other product and service names might be trademarks of IBM or other companies.

Notes:

Performance is in Internal Throughput Rate (ITR) ratio based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput that any user will experience will vary depending upon considerations such as the amount of multiprogramming in the user's job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve throughput improvements equivalent to the performance ratios stated here.

IBM hardware products are manufactured from new parts, or new and serviceable used parts. Regardless, our warranty terms apply.

All customer examples cited or described in this presentation are presented as illustrations of the manner in which some customers have used IBM products and the results they may have achieved. Actual environmental costs and performance characteristics will vary depending on individual customer configurations and conditions.

This publication was produced in the United States. IBM may not offer the products, services or features discussed in this document in other countries, and if offered, the terms may vary from those described. Consult your local IBM business contact for information on the product or services available in your area.

All statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.

Information about non-IBM products is obtained from the manufacturers of those products or their published announcements. IBM has not tested those products and cannot confirm the 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.

Prices subject to change without notice. Contact your IBM representative or Business Partner for the most current pricing in your geography.

This information provides only general descriptions of the types and portions of workloads that are eligible for execution on Specialty Engines (e.g. zIIPs, zAAPs, and IFLs) ("SEs"). IBM authorizes customers to use IBM SE only to execute the processing of Eligible Workloads of specific Programs expressly authorized by IBM as specified in the "Authorized Use Table for IBM Machines" provided at http://www.ibm.com/servers/eserver/serverprods/serverlibrary/autlookup.html ("AUT"). No other workload processing is authorized for execution on an SE. IBM offers SE at a lower price than General Processors/Central Processors because customers are authorized to use SEs only to process certain types and/or amounts of workloads as specified by IBM in the AUT.
Agenda

- Introduction
- Analytics Solutions
- What is Hadoop and why for Mainframe?
- Integration Scenarios
- DB2 for z/OS enabling Data Warehouse Consolidation and Data Flow Simplification
- Summary

... Enabling innovative insight with System z ...
System z Analytics Opportunity Based on Industry Trends

- Current analytics projects are mainly drawing on **z-based transactional data**
- **Real-time analytics-based decisions** is a top priority for most organizations
- These are advantages to the holder of operational data
- **Non-structured data** (LOGs, sensor data, social media, customer interaction) is of growing importance

**Types of data analyzed**

**Outperformers use analytics to drive better real time decisions**

- 77% Outperformer
- 58% Underperformer
- 33% more
Unfortunately for most of our Clients, the data Lifecycle is too fragmented to gain Advantage

- **Client key concerns:**
  - Cannot deliver *real-time analytics because of data latency* issues
  - Inadequate *performance*
  - Limited *governance model*
  - *Data completeness*
  - *Multiple platforms and systems*, many security boundaries, many points of failure, challenging recovery scenarios

- **Multi-day workshop**
  - *Captured the complexity of the current architecture*
The Analyst Community has taken Notice!

By eliminating analytic latency and data synchronization issues, hybrid transaction/analytical processing will enable IT leaders to simplify their information management infrastructure”

“This architecture will drive the most innovation in real-time analytics over the next 10 years via greater situation awareness and improved business agility”

Gartner Research Note G00259033:
Gartner 01-2014 Hybrid Transaction Analytical Processing Will Foster Opportunities
Business Analytics Solutions on zEnterprise

**Business analytics capabilities**

- **Cognos – Business Intelligence**
- **SPSS – Predictive Analytics**
- **TM1 – Performance Management**
- **BigInsights – Investigative Analytics**

**Business outcomes/benefits**

- Understand current & potential state
- Monitor results & fine-tune your business
- Inform strategy with a view into the future
- Predict customer segment & category affinity
- Market Basket Analysis to identify NBO
- Overlay browsing history onto purchase history to profile customers
- Reporting, analysis, operational & financial planning and consolidation
- Product profitability across customers, business & channels
- Sales Performance Management to improve efficiency in incentive compensation process
- Gain additional insights from LOGs, social media, streams, machine data, mass archives
- Understand and visualize the context of data in unstructured documents, LOGs and understand customer sentiment

© 2014 IBM Corporation
InfoSphere BigInsights for Linux on System z

- **What is it?**
  - A highly distributed parallel processing cluster for analyzing very large unstructured / differently structured data sets using a Hadoop and a Map/Reduce mode

- **What does it do?**
  - Breaks down large volumes of data into smaller chunks, processed in parallel across 100s/1000s of nodes and the results “stitched” back together
  - Provides an “Investigate analytics” approach
  - Generally a long running batch process to perform programmatic and investigative analysis
IBM InfoSphere BigInsights

Enterprise-grade Hadoop

100% open source Hadoop components

- Hadoop 2.2.0
- PIG 0.12.0
- Hive / HCatalog 0.12.0
- Oozie 3.3.2
- HBase 0.96
- Zookeeper 3.4.5
- Flume 1.3.1
- Avro 1.7.4
- Chukwa 0.5.0
- Solr / Lucene 4.7.2
- YARN 2.2.0*
- Spark 0.9.0*

Optional value-add components that do not compromise standards

- ANSI-compliant Big SQL – leverage common SQL syntax between mainframe DB2 and Hadoop dataset queries
- BigSheets – integrated spreadsheet based data manipulation and visualization for business users
- Text Analytics / AQL – development language and tooling for accelerated development of text analytic applications
- Advanced IDE – Comprehensive ECLIPSE IDE to accelerate development of Big Data applications
- Big-R
- Social media accelerators
- Log / Machine data accelerators

YARN = MapReduce 2.0 (MRv2)
Spark = Data analytics cluster computing framework
### InfoSphere BigInsights for Linux on System z

#### What's included?
- Enterprise Edition only offered on System z Linux
- Available from Fix Central August 5th, 2014
- GA release based on BigInsights 2.1.2
- Support RHEL 6.4
- Apache MapReduce and HDFS only (no GPFS FPO or Adaptive MapReduce)
- Pricing identical to System x and Power – per node/VM licensing, perpetual
- All components shown are entitled, some may need to be run on other platforms in this initial release

<table>
<thead>
<tr>
<th>Adaptive MapReduce and GPFS FPO are not supported in this release of BigInsights Enterprise Edition on System z.</th>
</tr>
</thead>
<tbody>
<tr>
<td>These components are included in the license but are separately installed and some components may need to be installed on platforms other than System z Linux.</td>
</tr>
<tr>
<td>Additional setup and integration is required. The open-source R language interpreter needs to be downloaded separately.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Big R</th>
</tr>
</thead>
<tbody>
<tr>
<td>Machine Data Analytics Accelerator</td>
</tr>
<tr>
<td>Social Media Analytics Accelerator</td>
</tr>
<tr>
<td>Text Analytics Accelerator (AQL)</td>
</tr>
<tr>
<td>Cognos bundle (business intelligence)</td>
</tr>
<tr>
<td>Streams (Real-time processing)</td>
</tr>
<tr>
<td>Data Explorer (Search &amp; Exploration)</td>
</tr>
<tr>
<td>IIG Integrations, Data Click</td>
</tr>
<tr>
<td>Adaptive MapReduce</td>
</tr>
<tr>
<td>IBM GPFS FPO</td>
</tr>
<tr>
<td>BigSheets</td>
</tr>
<tr>
<td>Big SQL</td>
</tr>
<tr>
<td>ECLIPSE IDE, developer tools</td>
</tr>
<tr>
<td>Web-based management console</td>
</tr>
<tr>
<td>Cluster installer</td>
</tr>
<tr>
<td>IBM Certified base Hadoop</td>
</tr>
</tbody>
</table>
Leverage Hadoop Innovation with System z Data
IBM InfoSphere BigInsights provides Options

- On-platform analysis of data that does not fit well with relational tools. Safely combine z/OS and external data for analysis
- Off-platform x86 and Power based Hadoop clusters can seamlessly take advantage of mainframe data

Customers looking into multiple Hadoop clusters, Data Lake, Marketing, Fraud, test, etc.
BigInsights for Linux on System z is another possible Hadoop cluster deployment
# Why BigInsights for Mainframe Customers?

<table>
<thead>
<tr>
<th>Business Reasons</th>
<th>Technical Reasons/Product Features</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Economies of scale</strong> – run the same Hadoop distribution both on and off the mainframe</td>
<td><strong>100% Hadoop compatible</strong> – Complete open source distribution with industry standard Hadoop tools</td>
</tr>
<tr>
<td><strong>Strategic flexibility</strong> – run the same Hadoop distribution across multiple platforms – System z, Power, System x, third party Intel-based HW</td>
<td><strong>Big SQL</strong> – leverage common SQL syntax for DB2 for z/OS and Hadoop queries</td>
</tr>
<tr>
<td><strong>Perpetual &amp; monthly</strong> license – flexible pricing options</td>
<td><strong>BigSheets</strong> – integrated spreadsheet based data manipulation and visualization for business users</td>
</tr>
<tr>
<td><strong>New skills</strong> – leverage new skills becoming readily available around the Hadoop ecosystem</td>
<td><strong>Text Analytics / AQL</strong> – development language and tooling for accelerated development of text analytic applications</td>
</tr>
<tr>
<td></td>
<td><strong>Advanced IDE</strong> – Comprehensive ECLIPSE IDE to accelerate the development of Big Data applications</td>
</tr>
</tbody>
</table>
Augmenting traditional EDW with Hadoop ...

<table>
<thead>
<tr>
<th>Aspects</th>
<th>Traditional EDW</th>
<th>BigInsights (Hadoop) enabled modernized EDW</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data Variety</strong></td>
<td><strong>Structured</strong> and <strong>semi-structured</strong> (e.g. XML) data, with pre-defined data modeling</td>
<td>Structured and <strong>unstructured</strong> data, schema-less (e.g. no fixed number of columns), data in any format, including LOGs, text</td>
</tr>
<tr>
<td><strong>Data Model</strong></td>
<td><strong>Schema-based</strong> relational data model, e.g. with star schemas, ROLAP/OLAP cubes, …</td>
<td>Analytics done on <strong>schema-less, non-modeled</strong> data that can determine what subset of data is valuable to the business</td>
</tr>
<tr>
<td><strong>Data Velocity</strong></td>
<td><strong>Batch</strong> oriented, nightly ETL to complete the processing before insight is obtained</td>
<td>Analytics is able to handle the data that moves fast, <strong>perform low-latency</strong>, real time analytics on any data formats</td>
</tr>
<tr>
<td><strong>Data Volume</strong></td>
<td><strong>Small</strong> to <strong>medium volume</strong> of data</td>
<td>Huge volume of data (e.g. <strong>Internet-scale</strong>), accommodating data growth rates of several TB/day</td>
</tr>
<tr>
<td><strong>Staging Area</strong></td>
<td>ETL based on <strong>predefined relational</strong> staging area with complex ETL</td>
<td>Hadoop augments the staging area to be used as an data ingestion and <strong>ETL pre-processing hub</strong></td>
</tr>
<tr>
<td><strong>Data Offload</strong></td>
<td>Stores data in transformed and aggregated structures, e.g. <strong>historical on IDAA</strong></td>
<td>Stores <strong>raw data</strong> to be stored as-is, providing off-load from EDW to affordable platforms</td>
</tr>
<tr>
<td><strong>Analytics Paradigm</strong></td>
<td><strong>‘Late’ analytical insight</strong> only after complex and costly ETL based on known analytical/reporting needs</td>
<td>Allows <strong>analytics, visualization, and exploration</strong> on raw data without executing costly ETL</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Big SQL</strong> allows Cognos BI based analytics on HDFS, HBase, Hive</td>
</tr>
<tr>
<td><strong>ETL / ELT</strong></td>
<td>Using <strong>ETL tools</strong>, such as InfoSphere DataStage on structured data</td>
<td>Augments conventional ETL with <strong>Streams-based</strong> real-time transformation and fully integrated Hadoop-based ETL/ELT</td>
</tr>
</tbody>
</table>
**Integrated Scenarios**

**Leveraging both Analytics Paradigms**

<table>
<thead>
<tr>
<th>Hadoop</th>
<th>SQL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scale-Out via parallel compute, I/O</td>
<td>Scale-Up</td>
</tr>
<tr>
<td>Key/value pairs (text docs, XML)</td>
<td>Relational tables</td>
</tr>
<tr>
<td>Functional programming</td>
<td>Declaritive queries</td>
</tr>
<tr>
<td>Offline batch (write once, read many)</td>
<td>Online transactions</td>
</tr>
</tbody>
</table>

---

**Developer Environments**

- Extract, Transform
- File Copy

**Business Intelligence Tools**

- Integrated Data Warehouse
- DB2 Analytics Accelerator for z/OS

---

- Sensor Data
- Blogs
- Web Data
- Docs
- PDFs
- Images
- Videos

- CRM
- SCM
- ERP
- Legacy
- 3rd Party
## Integrating Big Data Analytics with DB2 for z/OS

### Much of the world’s operational data resides on z/OS

- Two significant needs:
  - Merge this data with trusted OLTP and analytics data from zEnterprise data sources
  - Integrate this data so that insights from Big Data sources can drive business actions

### Unstructured data sources are growing

- Connectors to allow BigInsights to easily & efficiently access DB2 data
- DB2 is providing the connectors & the DB capability to allow DB2 apps to easily and efficiently access Hadoop data sources

---

### Diagram

- **IMS**
- **DB2**
- **IBM BigInsights**

- **Merge**
- **Integrate**

- **Relational projection of IMS model**

- **New user-defined functions and generic table UDF capability**
DB2 for z/OS v11: Integrating Big Data Sources

What is the insight derived through BigInsights?
- DB2 for z/OS submits a job to BigInsights using JSON Query Language (JAQL) via UDFs
  ➔ Here is a list of customer names/profiles, who are at risk leaving?
- JAQL job executes on BigInsights, stores the results on Hadoop HDFS
  ➔ These are customers at risk of leaving
- DB2 for z/OS executes a table UDF HDFS_READ to retrieve data from HDFS
- DB2 for z/OS executes JOINs with relevant customer profile data and other relevant tables to complete the analysis
  ➔ Execute Next Best Action via Campaign Management for customers at risk
Hadoop on the Mainframe or rather off? What makes sense when?

Case 1: Hadoop on the Mainframe
- Data originates mostly on the mainframe (Log files, database extracts)
- Data security a primary concern
- Clients will not send data across external net
- Relatively small data – 100 GB to 10s of TBs
- Hadoop is valued mainly for richness of tools
- Z governance and security models needed

Case 2: Hadoop off the Mainframe
- Much data originates off the mainframe
- Security less of a concern since data is not trusted anyway
- Very large data sets – 100s of TB to Pbs
- Hadoop is valued for ability to manage large datasets economically
- Desire to leverage cheap processing and potentially cloud elasticity
IBM DB2 Analytics Accelerator

**What is it?**
- A high performance appliance that integrates Netezza technology with zEnterprise technology, to deliver dramatically faster business analytics on relational / structured data

**What does it do?**
- Accelerates complex SQL queries, up to 2000x faster (days, hours reduced to seconds)
- Improves access to and lowers the cost of storing, managing and processing historical data
- Minimizes latency
- Reduces zEnterprise capacity requirements
- Improves security & reduces risk
Operations and Analytics Coexistence

Two use cases:

- **Operational Priority**
  - Keeping operational throughput constant, add analytics load to the system
  - Data used for analysis can be slightly out of sync with operations

- **Data Priority**
  - Data used for operations and analytics must be in close synchronization
  - Higher latency of operational throughput is acceptable

OLTP Transactions
- Real-time data ingestion
- High concurrency
- Advanced analytics
- Standard reports

OLAP

Operational analytics
- Complex queries
- Historical queries

Advanced analytics

Operational analytics

Complex queries

Historical queries

Data Priority

Coupling Facility

OLAP

Operational analytics

Complex queries

Historical queries

OLTP Transactions

Real-time data ingestion

High concurrency

Advanced analytics

Standard reports

OLAP

Operational analytics

Complex queries

Historical queries

z/OS LPAR Configuration: zEC12, 8 CPUs; 120 GB Memory; zOS 1.13; DB2 V11 for z/OS
IBM DB2 Analytics Accelerator Configuration: Striper model, 112 cores; 240 data slices
IBM DB2 Analytics Accelerator Configuration updated via batch processes such as nightly ETL
## EDW Modernization

**Sample “Client“ Implementation Steps**

<table>
<thead>
<tr>
<th>Today's Environment</th>
<th>Intermediate Steps</th>
<th>Target Architecture</th>
<th>Target Architecture with Hadoop</th>
<th>Target Architecture with Hadoop with MDM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Traditional multi-layered EDW (FTP/staging, SOR, data marts)</td>
<td><strong>Enabling ODS</strong> through DB2 for z/OS with IDAA</td>
<td>Enabling Hybrid Transactional/Analytical Processing (HTAP) enabled by DB2 for z/OS with IDAA performance capabilities at operational and EDW level</td>
<td>Introducing <strong>Hadoop</strong> via BigInsights for text and social media analytics</td>
<td>Implementing <strong>MDM</strong> with operational integration</td>
</tr>
<tr>
<td>Time consuming t+3 information supply chain to build data mart layer</td>
<td>Simplification of entire information supply chain, including ELT through DataStage Balanced Optimization</td>
<td>Data mart elimination (for performance reasons)</td>
<td>Bi-directional integration with EDW for pervasive analytical insight</td>
<td>Pervasive Master Data governance for core enterprise information (e.g. party)</td>
</tr>
<tr>
<td><strong>Lack of ODS</strong></td>
<td>Initial <strong>MDM</strong> value for single view</td>
<td>Leveraging Hadoop ELT via DataStage</td>
<td>Bi-directional integration with zEDW for analytical insight</td>
<td></td>
</tr>
<tr>
<td><strong>No HTAP capabilities</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
DW/BI Modernization – Initial Consideration

- **Integration with Hadoop-based Systems**
  - Enrichment of structured data with semi- and non-structured data
  - Data Exploration
  - Analytics on semi- and non-structured data, e.g. Log analytics, XML, …

- **Provide analytical capabilities** to the business in (near) real-time
  - Descriptive (query drill down, ad-hoc queries, reporting)
  - Predictive (modeling)
  - Prescriptive (scoring, adaptive business rules)

- **DB2/IDAA** – Hybrid Transactional and Analytical Processing (HTAP) on business critical, structured data

- **Enhancements for optimized processing of historical data on DB2/IDAA**
  - DB2 10 Bi-temporal
  - DB2 11 transparent archiving
  - IDAA High Performance Storage Saver
Vielen Dank