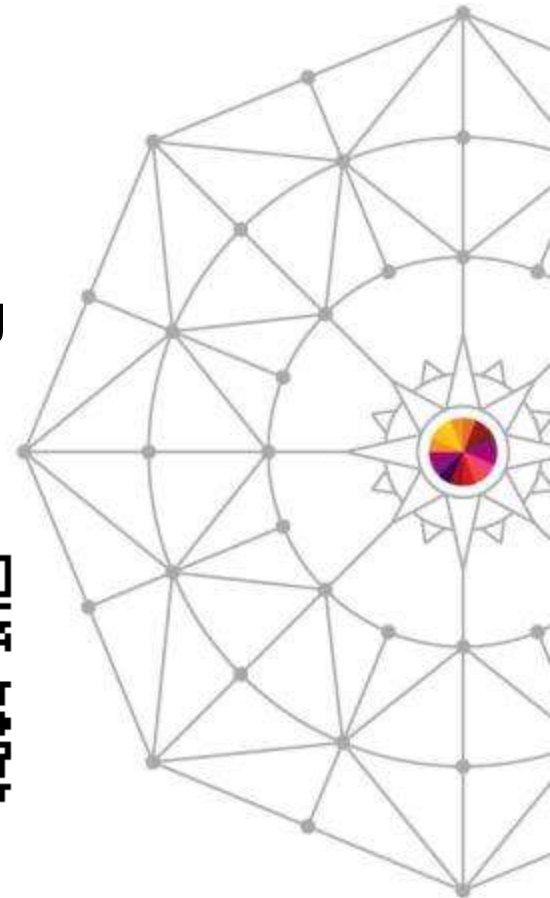# Big Data Strategies with DB2 for z/OS

Mark Simmonds – WW Product Marketing and IT architect

Joachim Limburg – Director DB2 for z/OS

IBM Corporation

March 13, 2014
Session Number 14729

Please note

IBM's statements regarding its plans, directions, and intent are subject to change or withdrawal without notice at IBM's sole discretion.

Information regarding potential future products is intended to outline our general product direction and it should not be relied on in making a purchasing decision.

The information mentioned regarding potential future products is not a commitment, promise, or legal obligation to deliver any material, code or functionality. Information about potential future products may not be incorporated into any contract. The development, release, and timing of any future features or functionality described for our products remains at our sole discretion.

Performance is based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput or performance that any user will experience will vary depending upon many factors, including 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 results similar to those stated here.

Acknowledgements and Disclaimers

Agenda

- Big Data in an Information Driven economy
- Why start with System z
- DB2 for z/OS and InfoSphere BigInsights /Hadoop
- Summary / Call to action

On a Smarter Planet, Unprecedented Changes are Occurring

ZBs of transaction data

Cloud

Mobile

Big Data

Social

Internet of Things

- Business models under constant pressure

- Customers are more demanding and connected

- Great relationships trump great products

Complete your session evaluations online at www.SHARE.org/Anaheim-Eval

Technology is Transforming Business and IT Collaboration

Factors Impacting Organizations

2004	2006	2008	2010	2012

1. Technology Factors

1. Technology Factors

2. People Skills

3. Market Factors

4. Macro-economic Factors

5. Regulatory Concerns

6. Globalization

Source: IBM CEO Study 2012

And Leaders are Responding by…

**Providing a
Great Experience**

**Offering Value
In Every
Interaction**

**Innovating
Across
the Ecosystem**

Complete your session evaluations online at www.SHARE.org/Anaheim-Eval

But what is Big data?

> *Google can give you nearly 2 Billion options*
> *Vendors have even more definitions*

What is Big Data

Web   Images   Maps   Shopping   News

About 1,940,000,000 results (0.18 seconds)
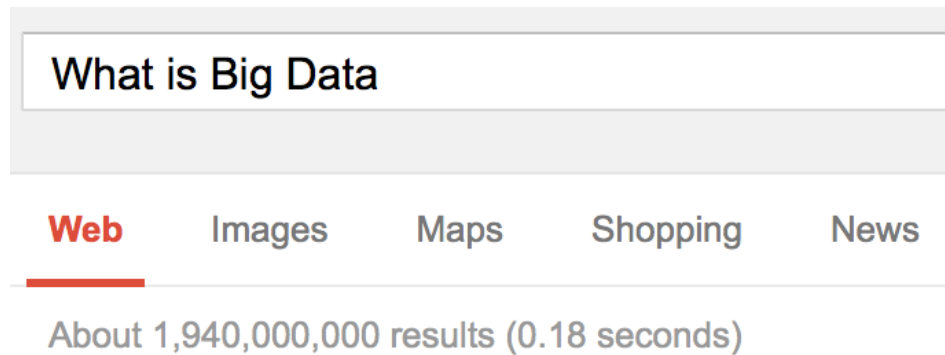
Here is how Gartner defines Big Data

> *Big data is high-volume, high-velocity and high-variety information assets that demand* **cost-effective, innovative information processing for enhanced insight and decision making**.

Gartner research note "Survey Analysis - Big Data Adoption in 2013 Shows Substance Behind the Hype" Sept 12 2013
Analyst(s): Lisa Kart, Nick Heudecker, Frank Buytendijk
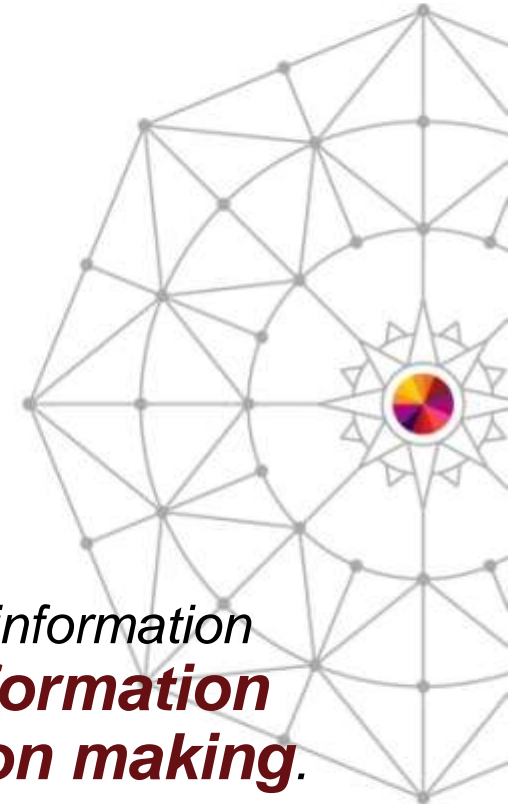
We've Moved into a New Era of Computing - V^4

Radical Flexibility

12 terabytes

of Tweets create daily
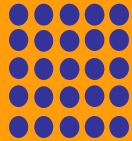
Extreme Scalability

5 million

trade events per second
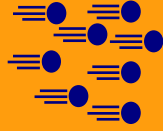
Volume   **Velocity**   **Variety**   **Veracity**

100's

Of video feeds from surveillance cameras

Information from everywhere

Only **1 in 3**

Decision makers trust their information

"We have for the first time an economy based on a key resource [Information] that is not only renewable, but self-generating.

Running out of it is not a problem, *but drowning in it is*."

– John Naisbitt

Agenda

- Big Data in an Information Driven economy
- Why start with System z
- DB2 for z/OS and InfoSphere BigInsights /Hadoop
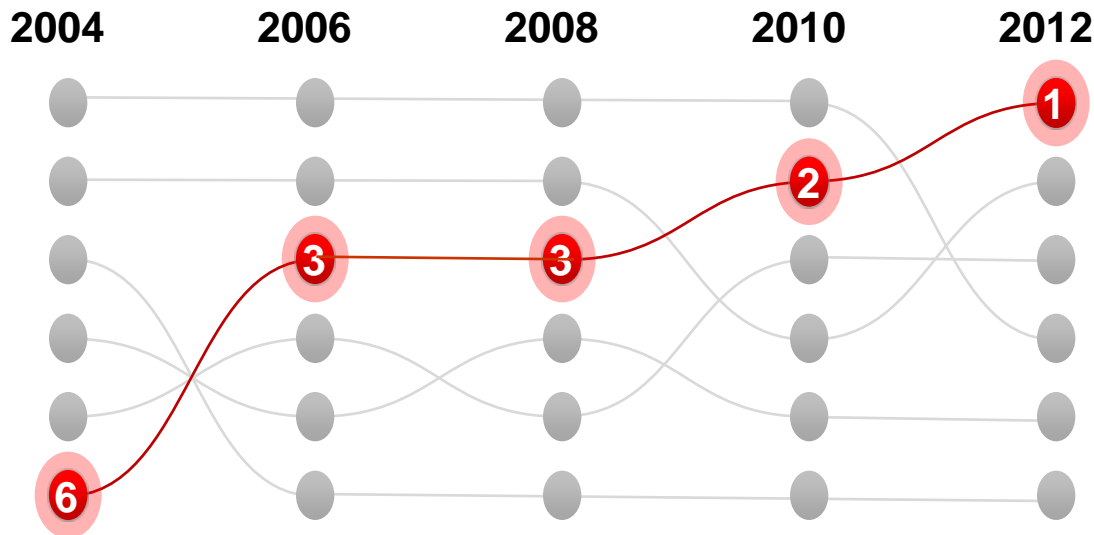- Summary / Call to action
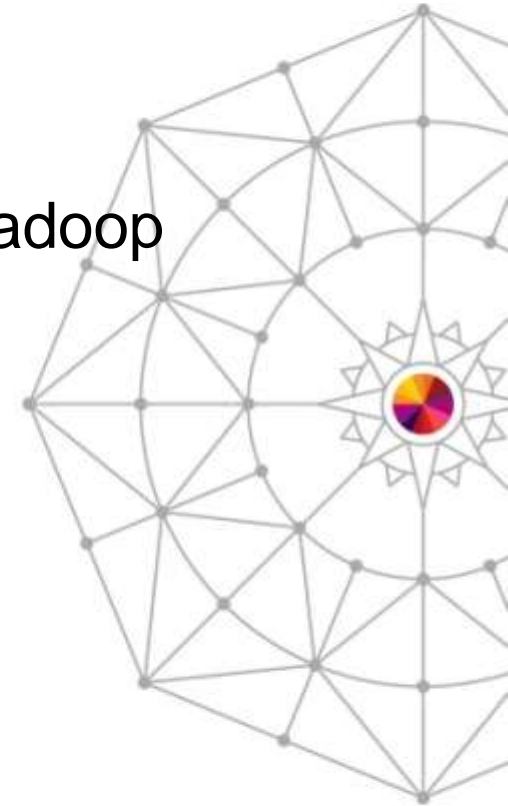
The Big Data Starting Point

Types of Data Analysed

Transactional sources are the dominant data types analyzed in big data initiatives
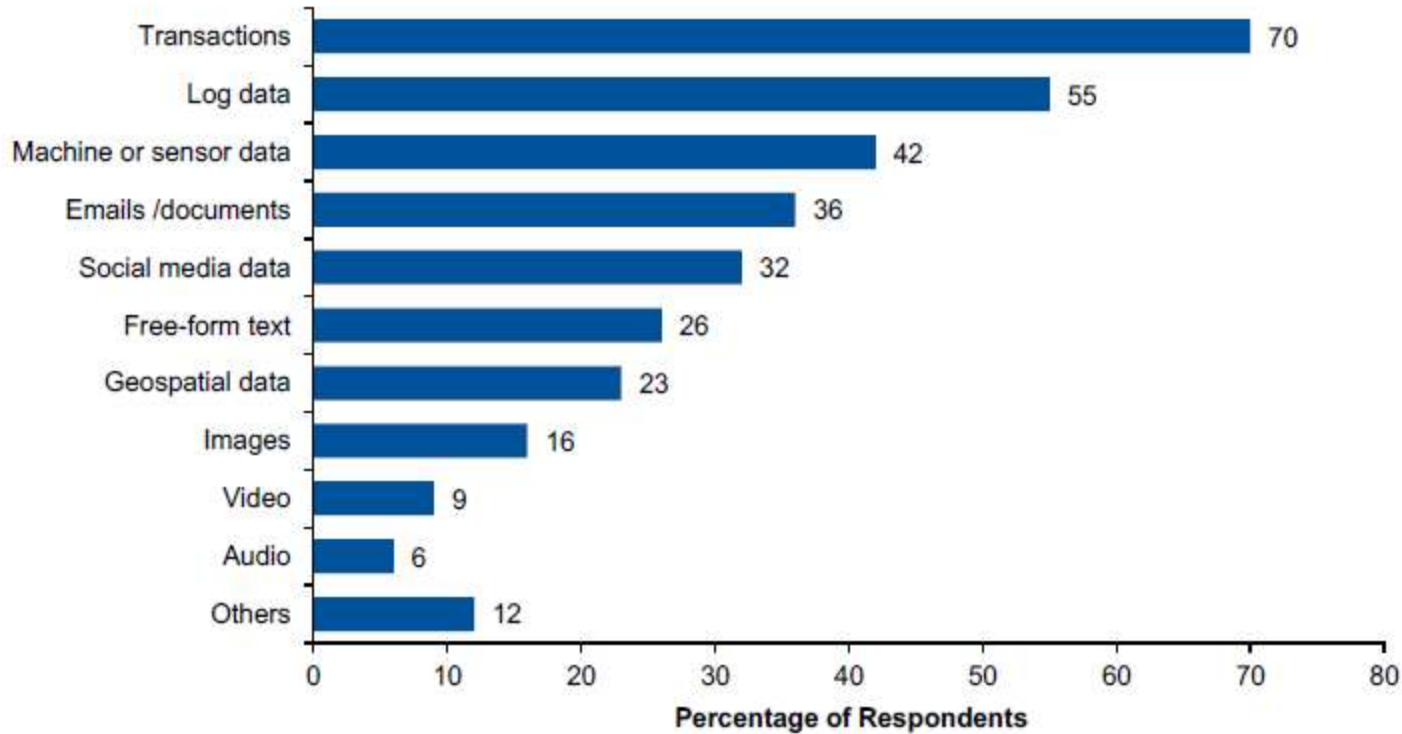
N =465 (multiple responses allowed)

Source: Gartner (September 2013)

Gartner research note "Survey Analysis - Big Data Adoption in 2013 Shows Substance Behind the Hype" Sept 12 2013
Analyst(s): Lisa Kart, Nick Heudecker, Frank Buytendijk

The Big Data Starting Point

Types of Big Data Analysed by Industry

Transactional sources are the dominant data types analyzed in big data initiatives

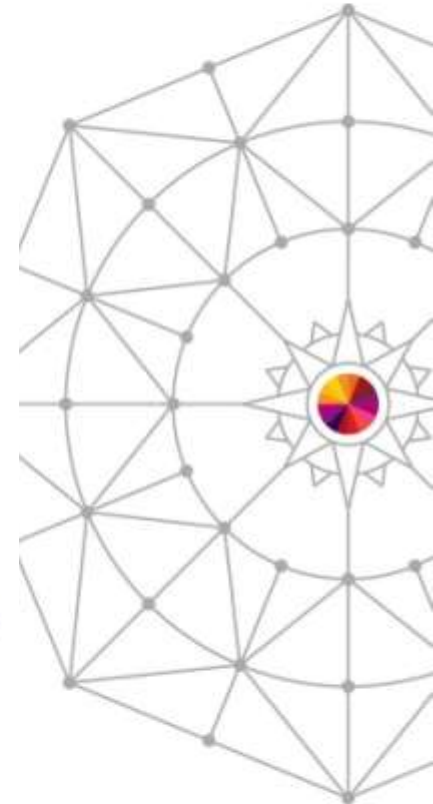|	Manufacturing and Natural Resources	Media/ Communications	Services	Government	Education	Retail	Banking	Insurance	Healthcare	Transportation	Utilities
Transactions	73%	62%	67%	67%	54%	93%	83%	81%	75%	79%	80%
Log data	44%	57%	58%	59%	54%	40%	66%	61%	33%	71%	60%
Machine or sensor data	53%	38%	35%	33%	31%	27%	27%	48%	42%	50%	40%
Emails /documents	27%	43%	43%	41%	46%	27%	34%	39%	17%	29%	20%
Social media data	32%	52%	39%	26%	54%	73%	27%	13%	-	50%	-
Free-form text	17%	24%	28%	30%	31%	20%	34%	35%	67%	21%	40%
Geospatial data	27%	14%	19%	19%	38%	27%	27%	26%	8%	29%	40%
Images	19%	24%	17%	11%	38%	13%	5%	16%	25%	7%	-
Video	8%	29%	12%	7%	31%	13%	-	6%	8%	7%	-
Audio	10%	19%	8%	4%	8%	-	-	6%	-	-	-
Other	8%	14%	13%	15%	8%	7%	10%	16%	42%	14%	-
n =	59	21*	127	27*	13*	15*	41	31	12*	14*	5*

Note: Highlighted cells indicate the top three data types by industry.
Multiple responses allowed

Complete your session evaluations online at www.SHARE.org/Anaheim-Eval

…and our webcast survey said…

- Have you already implemented or are you planning to implement any Big Data based initiatives within the next 6 months?

  Yes 31%
  No 69%

- How would you rate the value of being able to integrate insights from social media, telemetry, unstructured data into your analytics and decision making processes?

  High 50%
  Med 36%
  Low 14%

- Do you see the IBM System z platform as ~~~ success of Big Data ~~~

  Yes 90%
  No 10%

Five key findings from the study about big data:
1. Customer analytics are driving big data initiatives
2. Big data is dependent upon a scalable and extensible information foundation
3. Initial big data efforts are focused on gaining insights from internal data
4. Big data requires strong analytics capabilities
5. Adoption of big data is focused upon delivering measureable business value, which happens in four stages:
   - Educate: focusing on business as usual with casual understanding of big data;
   - Explore: developing strategy and roadmap based on business needs and challenges;
   - Engage: creating pilots to validate value and requirements; and
   - Execute: deploying two or more big data technologies and continuing to innovate

www.SHARE.org/Anaheim-Eval

Said Business School at the University of Oxford

Value (IBV,) with the
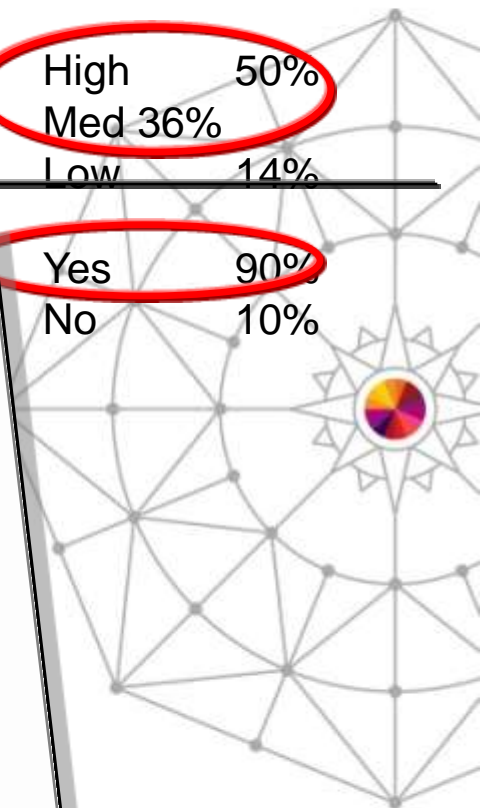
The role of zEnterprise in Big Data

- A large percent of the data that is accessed for analytics originates/resides on IBM zEnterprise
  - 2/3 of business transactions for U.S. retail banks
  - 80% of world's corporate data
- Businesses that run on zEnterprise
  - 66 of the top 66 worldwide banks
  - 24 of the top 25 U.S. retailers
  - 10 of the top 10 global life/health insurance providers
- 1,300+ ISVs run zEnterprise today, more than 275 of these selling over 800 applications on Linux
- The downtime of an application running on System z equates to approximately 5 minutes per year
- The System z mainframe can run over a thousand virtual Linux images on a single frame the size of a refrigerator

Majority of today's analytics based on relational / "Structured" Data

- Analytics and decision engines reside where the DWH / transaction data is

- "Noise" (veracity) surrounds the core business data

  - Social Media, emails, docs, telemetry, voice, video, content

- What data are you prepared to *TRUST?*

- Where do you put your trusted Data?

"Circle of trust"

Data Warehouse

Integration

Business Analytics

DB2 for z/OS IMS

Information Governance

Demand for _differently_ structured data to be seamlessly integrated, to augment analytics / decisions

- Analytics and decision engines reside where the DWH / transaction data is

- "Noise" (veractity) surrounds the core business data

    - Social Media, emails, docs, telemetry, voice, video, content

- Expanding our insights – getting closer to the "truth"

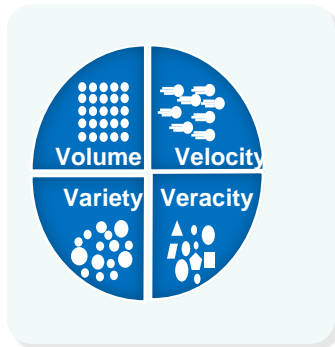   - Lower risk and cost

    - Increased profitability

"Circle of trust" widens

Data Warehouse

Integration

Business Analytics

DB2 for z/OS IMS

Information Governance

Forward Thinking Organizations are Creating Value From Big Data

The power of Data coming together…

…to deliver Improved Business Outcomes

…with the power of Technology…

1. **Enrich your information base**
   with Big Data Exploration

2. **Improve customer interaction**
   with Enhanced 360º View of the Customer

3. **Optimize operations**
   with Operations Analysis

4. **Gain IT efficiency and scale**
   with Data Warehouse Augmentation

5. **Prevent crime**
   with Security and Intelligence Extension

Fraud Detection – Claiming disability allowance.

"Unable to work"

Work Status

"Dude – awesome vacation"

Facebook Post

Make payment or investigate
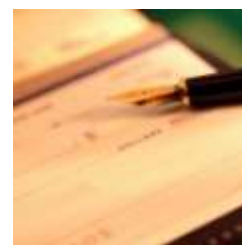
Deterrent for fraudsters - Cost Savings for the business

zEnterprise

Data from Social Media sites analyzed with Text analytics
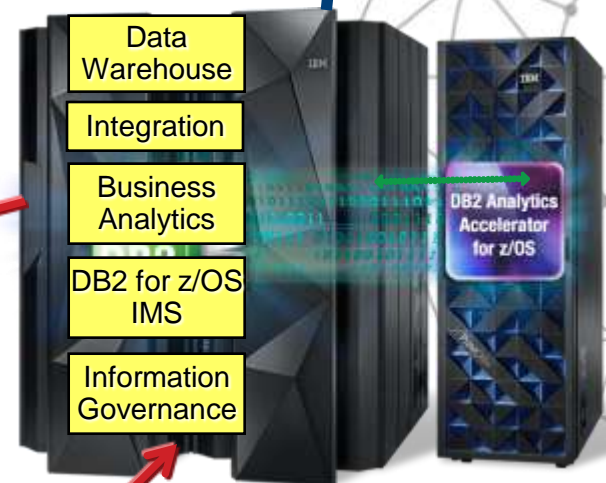
Refined Search parameters from OLTP environment

Hadoop or agency

| Data Warehouse |
| Integration |
| Business Analytics |
| DB2 for z/OS IMS |
| Information Governance |

DB2 Analytics Accelerator for z/OS

Result Set for further processing
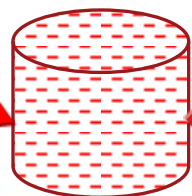
Data Warehouse + modeling applications

Result set uploaded or directly imported into OLTAP DBMS

Enterprise Integration and Governance…
the key to success of incorporating Big Data

- **Information Integration**

  - Insights from big data must be incorporated into the warehouse and analytics/ decision engines

- **Information Governance**

  - Companies need to govern what comes in, and the insights that come out

Data Warehouse

Big Data Platform

Enterprise Integration

Traditional Sources

New Sources

IBM BIG DATA PLATFORM
Logical platform with many physical deployment options

BIG DATA PLATFORM

Systems Management | Application Development | Discovery

Accelerators

Hadoop System | Stream Computing | Data Warehouse

Information Integration & Governance
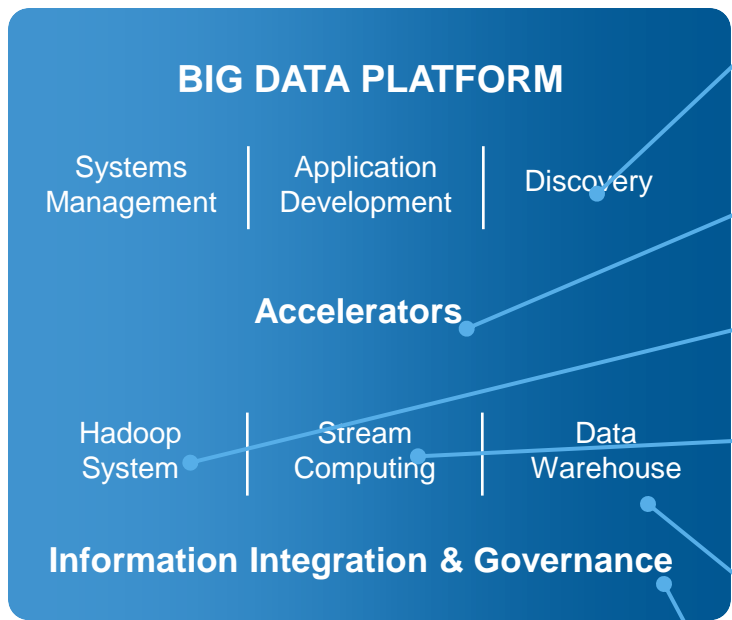
InfoSphere Data Explorer
Find, navigate, visualize all data

Accelerators
Speed time to value with analytic and application accelerators

InfoSphere BigInsights
Bringing Hadoop to the enterprise

InfoSphere Streams
Analytics for data in-motion exploration

InfoSphere Data Warehouse
Delivers deep insight with advanced database analytics & operational analytics

Information Integration and Governance
Governs data quality and manages the information lifecycle
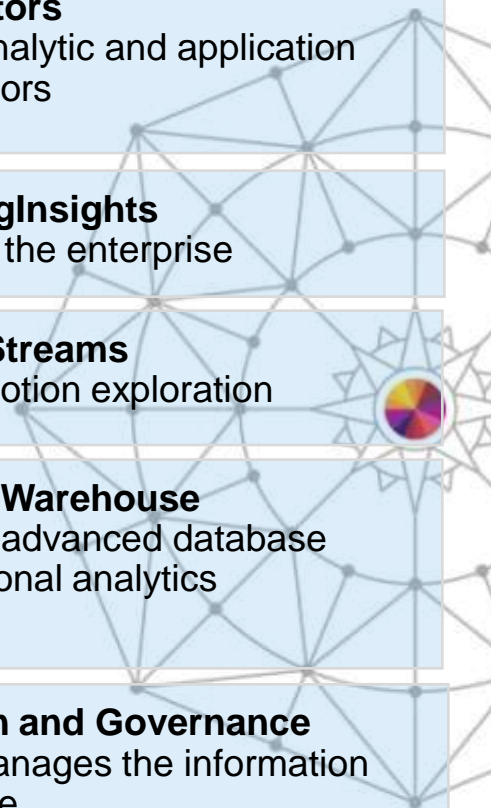
IBM DB2 Analytics Accelerator v 4.1

Do things you could never do before!

- ## What is it?

  - A high performance appliance that integrates Netezza technology with zEnterprise technology, to deliver dramatically faster business analytics

- ## What does it do?

  - Accelerates complex queries, up to 2000x faster

  - Lowers the cost of storing, managing and processing historical data

  - Minimizes latency

  - Improves security and reduces risk

  - Complements existing investments

IBM PureData System for Hadoop

Accelerate Hadoop analytics with appliance simplicity

Accelerate Big Data projects with built-in expertise
- Explore new ways to use all data
- Unlock new insights from unstructured data
- Establish a cost efficient on-line data archive

Simplify with integrated system management
- InfoSphere BigInsights software
- Compute and Storage hardware

Ensure production grade security and governance
Easily integrate with other systems
in the IBM big data platform

Simplify Big Data for the enterprise!
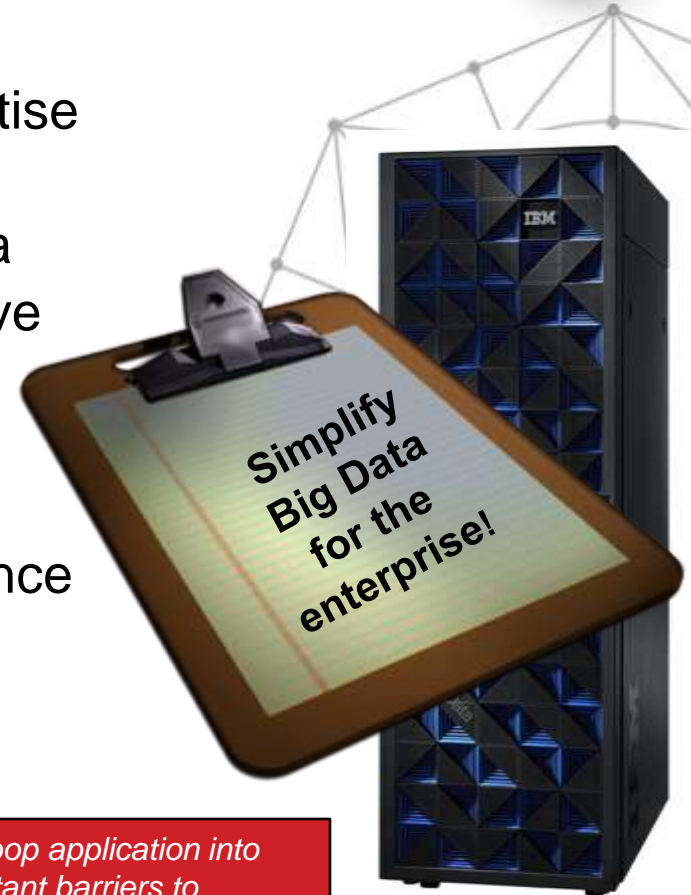
CIO:Insight Apr 29 2013…Issues surrounding how long it takes to get a Hadoop application into production coupled with a lack of real-time capabilities are proving to be important barriers to deployment. As a result, the respondents are reporting that both the number of Hadoop applications and the size of the overall Hadoop environment remain relatively small.

New

SHARE
in Anaheim

Approaches

1

Processing done outside z
(Extract and move data)

Move data over network

Log files in z/OS

$$$$
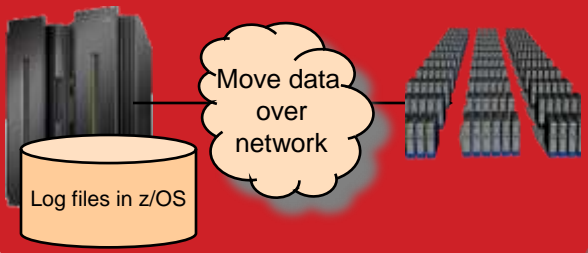
Additional infrastructure.
Challenges with scale,
governance, ingestion.

2

Processing in Appliance
(z remains the master)

MR jobs, result

Log files in z/OS
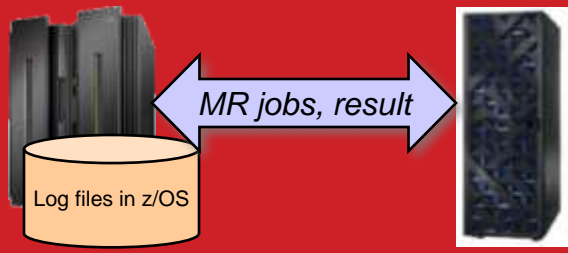
$$

"IDAA like" approach with
PureData System for
Hadoop.
High speed load.
z is the control point.

3

Processing done on
System z
(MR cluster on zLinux)

DB2

Log files in z/OS   z/OS   z/VM

$$$
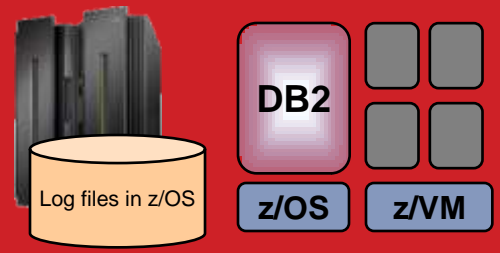
Provision new node quickly
Near linear scale.
High speed load.
z is the control point.

Data Explorer : visualization & discovery across all your data sources : "*Integration at the glass*"

InfoSphere Data Explorer

Providing unified, real-time access and fusion of big data unlocks greater insight and ROI

Securely connect to and leverage data stored in DB2 for z/OS & IMS

Help prioritize your System z big data integration and analytics projects

Improve customer service & reduce call times

Create unified view of ALL information for real-time monitoring

Increase productivity & leverage past work increasing speed to market

Analyze customer information & data to unlock true customer value

Identify areas of information risk & ensure data compliance

SHARE in Anaheim

Agenda

- Big Data in an Information Driven economy
- Why start with System z
- DB2 for z/OS and InfoSphere BigInsights /Hadoop
- Summary / Call to action

What does a Big Data platform do?

Analyze a Variety of Information

Novel analytics on a broad set of mixed information that could not be analyzed before

Analyze Information in Motion

Streaming data analysis

Large volume data bursts & ad-hoc analysis

Analyze Extreme Volumes of Information

Cost-efficiently process and analyze petabytes of information

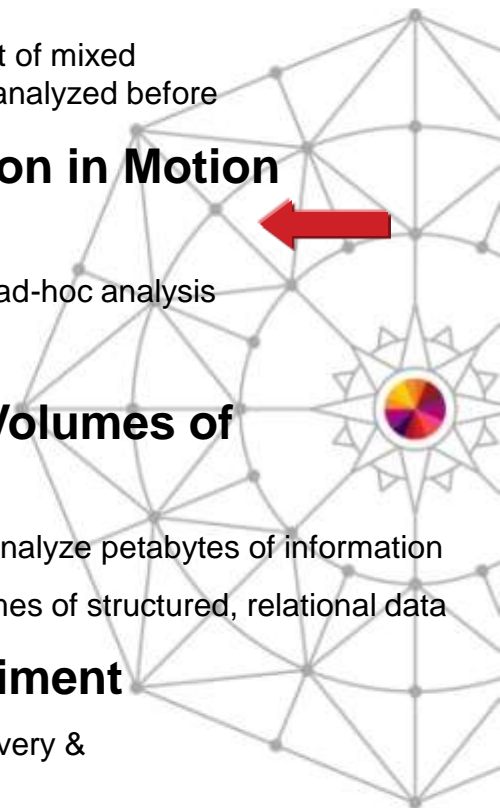Manage & analyze high volumes of structured, relational data

Discover & Experiment

Ad-hoc analytics, data discovery & experimentation

Manage & Plan

Enforce data structure, integrity and control to ensure consistency for repeatable queries

$\mathcal{H}$

How is Big Data Managed

Hadoop Distributed File System (HDFS)

HDFS

Requires Linux OS

•Write once, read many
•Optimized for streaming in and out
•Each file block is replicated, typically three times
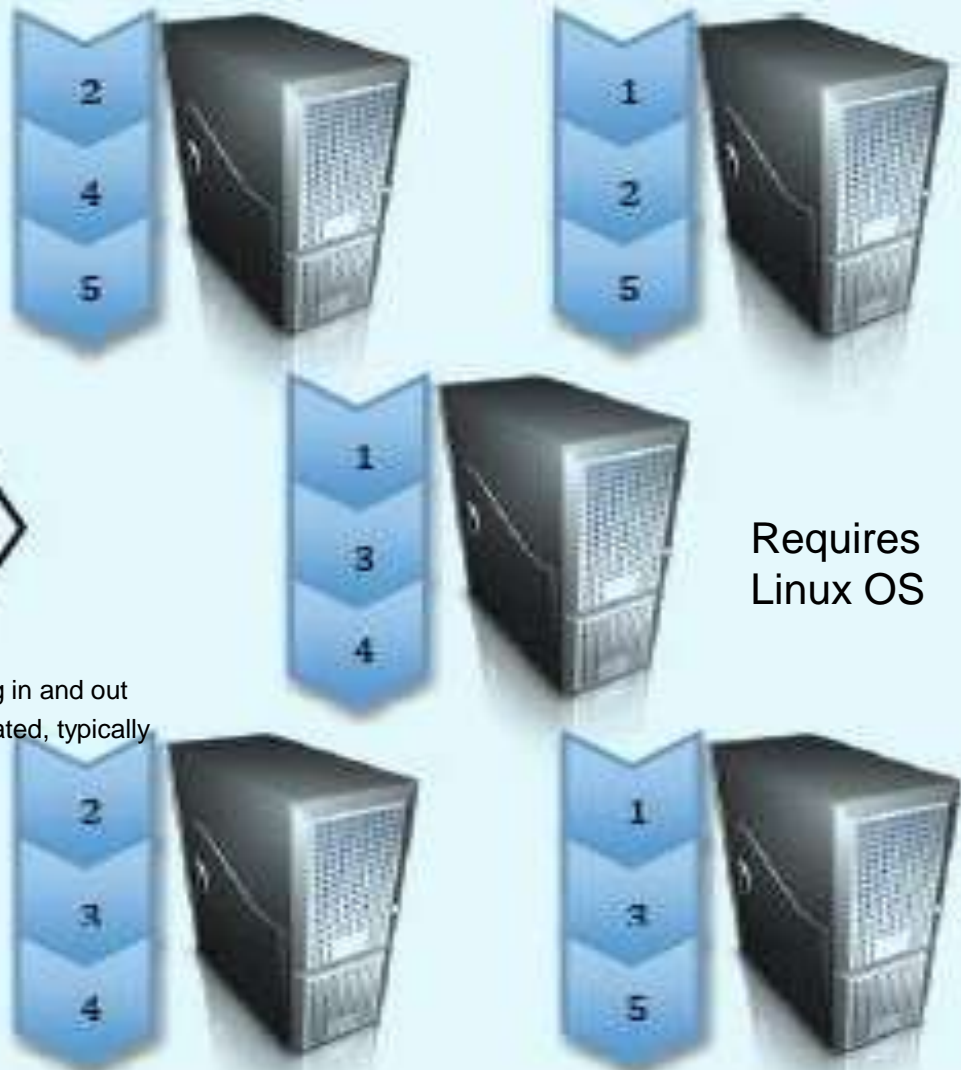
 HDFS breaks incoming files into blocks and stores them redundantly across the cluster.

MapReduce

- A simple, yet powerful framework for parallel computation
    - Applicable to many problems, flexible data format
- Basic steps:

    - Do parallel computation (Map) on each block (split) of data in an HDFS file and output a stream of (key, value) pairs to the local file system
    - Redistribute (shuffle) the map output by key
    - Do another parallel computation on the redistributed map output and write results into HDFS (Reduce)

Mapper   Reducer

M_1 M_2 M_3 M_4 shuffle R_1 R_2
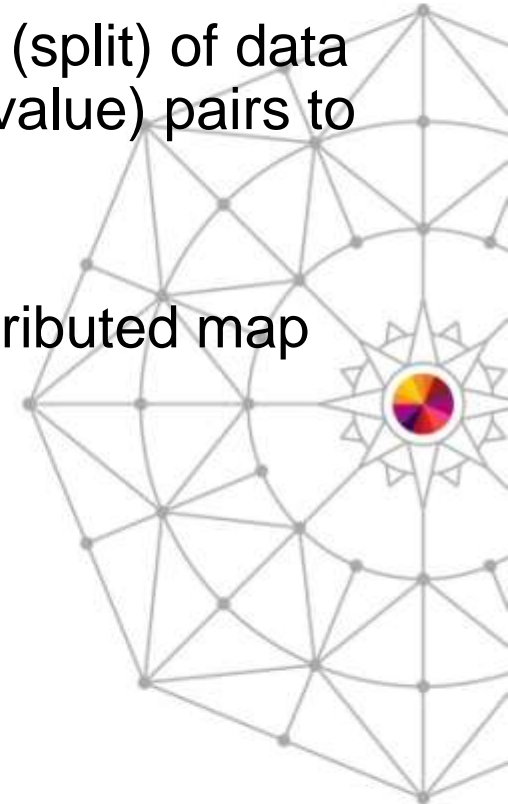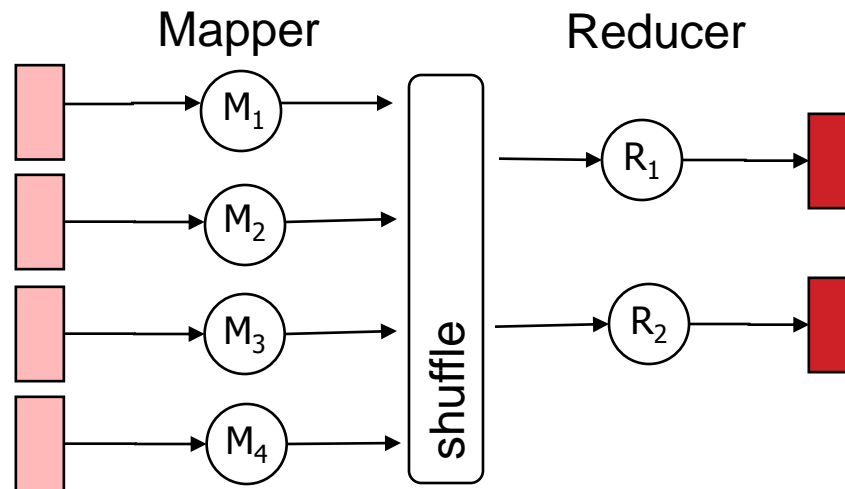
Managing big data volumes is only half the battle

- Real benefit derived from big data integration with traditional corporate data
  - Fitting the insight gained from analyzing these new types of information into their existing business processes and operations.

- Database vendors now integrating Hadoop into their products
  - DB2 LUW, Oracle, SQL server, HP, Teradata
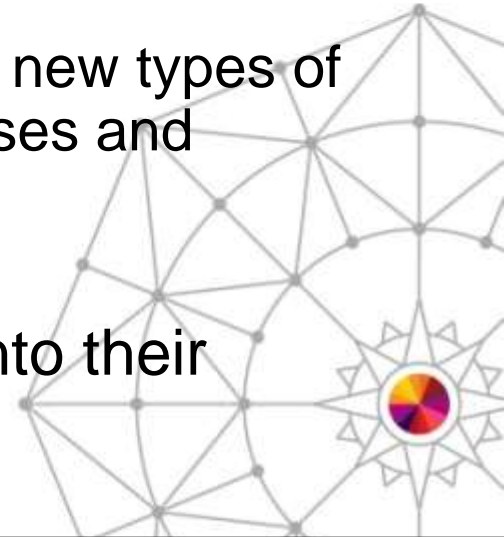
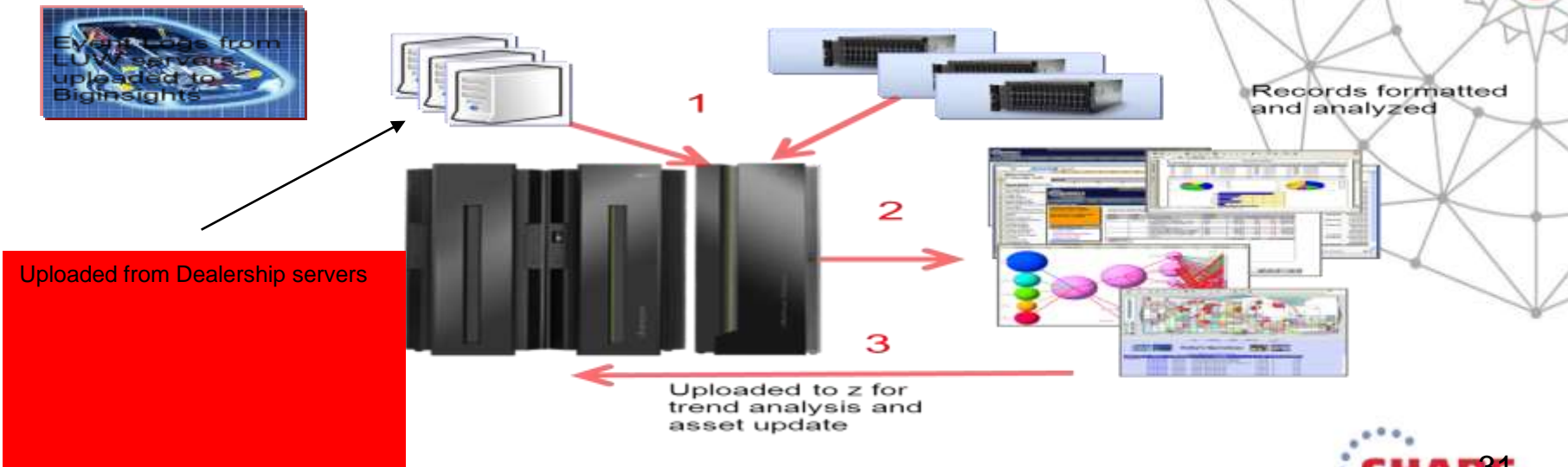| Real-time Traffic Flow Optimization | Precise fraud & risk detection | Understand and act on customer sentiment | Accurate and timely threat detection | Predict and act on intent to purchase | Low-latency network analysis |

A use case to access BigInsights from DB2

- BigInsights ingests data that usually is not ingested by established structured data analysis systems like DB2 for z/OS ,.e.g., email from all clients sent to an insurance company

- <u>DB2 kicks off a Hadoop job on BigInsights</u> that analyzes the emails and identifies customers who have expressed dissatisfaction with our service and the word 'cancel', 'terminate', 'switch' or synonyms thereof, or names of our competitors.

- <u>BigInsights job</u> runs successfully, <u>creates a file of results</u> (names and email addresses of customers at risk) and terminates.

- <u>DB2 reads the BigInsights result file</u> using it's "user defined function" methodology.

- DB2 joins the result with the Agent table and alerts the agents of the at-risk customers. The agents upon the at-risk customer and offer a promotion to stave off defection.

30
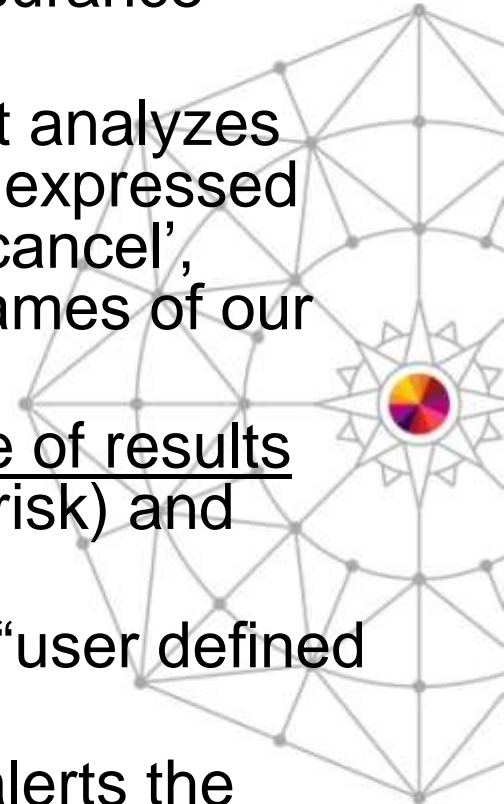
Scenario 1: Automobile manufacturer

Challenge / Customer pain point	Current technologies	Bigdata/Biginsights
Manufacturers need to be proactive in responding to potential quality and safety issues with vehicles **The Sheer number and differing formats of these logs makes it impossible for customers to investigate and identify potential issues or failures in a pre-emptive way** **Almost impossible.**	– Tracking of claims and Warranty repairs coming in from dealerships – On board computer data is used by dealerships and only uploaded to manufacturer when requested.	Solution: **1.Automatically upload Onboard vehicle data to Hadoop.** **2.Uses BigInsights tooling and templates to establish baseline parameters and queries to identify potential problems or failure trends** **3.Upload result set to z for integration with Warranty claims details to identify trends** Benefits: • **Proactive response to potential issues** • **Cost reductions with fewer warranty claims before identification..**

Uploaded from Dealership servers

Records formatted and analyzed

1

2

3

Uploaded to z for trend analysis and asset update

31

Generic table UDF

- Existing DB2 TUDF requires the table schema to specified at create time

  - Different BigInsight Analytic tasks return different number and types of columns.

  - This approach could require the user to create up to one table udf per analytic function defined on zDB2 but run on BigInsights

- **Generic TUDF does not specify output table schema at create time but at reference time**

  - The same table udf may be invoked by different analytic functions defined on zDB2, the schema on the DB2 is specified at the time of query

  - Inability to handle analytic algorithms where the schema is known only at execution time, e.g., dimension reductions, where the reduced number of dimensions are known only after the analytics is run.
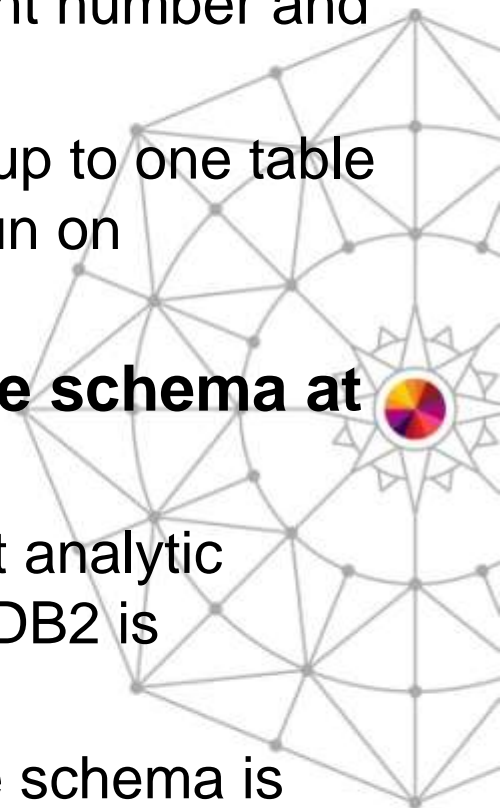
Example of generic table function

```
CREATE FUNCTION  hdfs_Read (handle  VARCHAR (1000) )
 RETURNS GENERIC TABLE
 LANGUAGE C PARAMETER STYLE SQL… ;


SELECT TX.*
FROM TABLE ( hdfs_Read( 'http://172.16.134.134:8080/data/sample.csv ')
           AS TX (Seller  varchar(20),
                  Year    INT,
                  Total    decimal(15,2));


SELECT TX.*
FROM TABLE ( hdfs_Read( 'http://172.16.134.134:8080/data/weather.csv ')
      AS TX (City  varchar(20),
             High_temperature    Decimal(5,2));
```
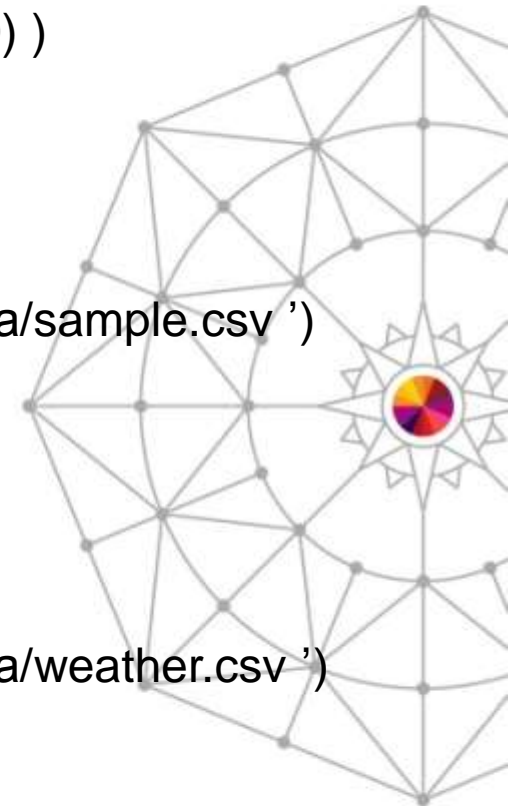
HDFS_Read

- It is a generic table UDF packaged with bigInsight
- Connect IBM BigInsights platform via REST API.
- Read a file from HDFS in delimited form and returns a table based on the schema specified in the query.
- HDFS_READ will use the BigInsights REST API for the data transfer.

- HDFS_READ function will take two arguments:
  - >>-HDFS_READ—(--FILE-URL--,-option-string-)----------------
    --------><
  - URL: Server address and path of the file in the HDFS
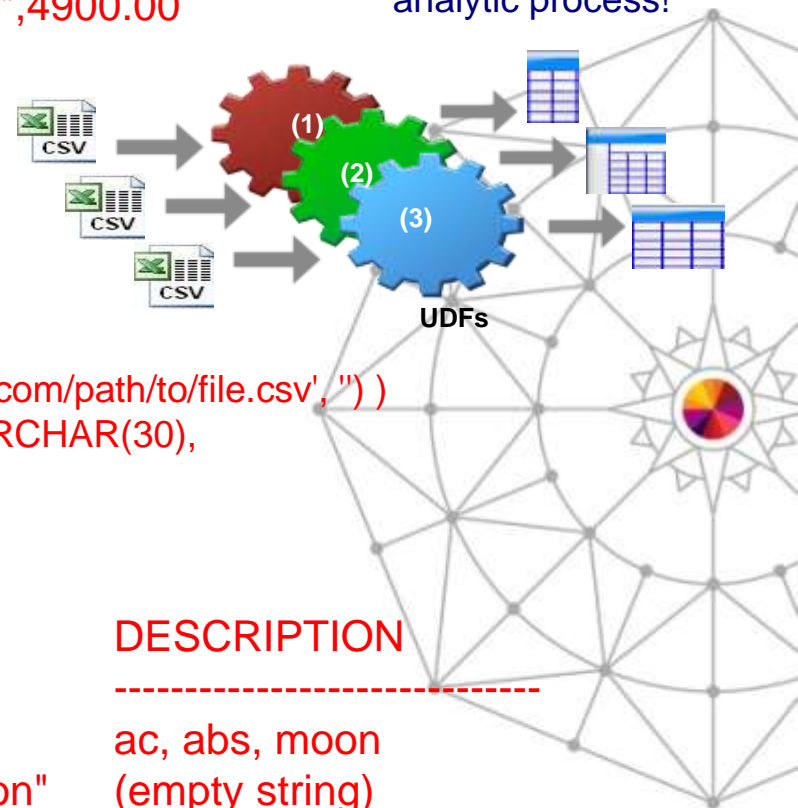 - Option-string: specify "DELIMITER"
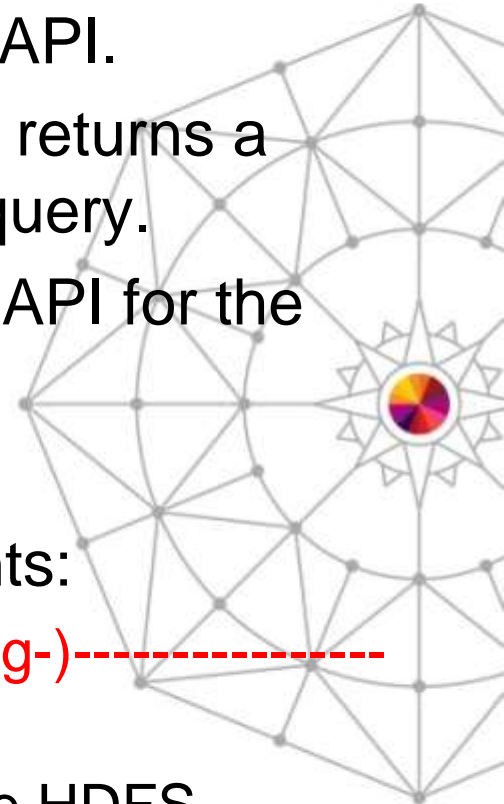
Example CSV file with SQL

- The following is an example of a CSV file.

  1997,Ford, E350,"ac, abs, moon",3000.00
  1999,Chevy, "Venture ""Extended Edition""","",4900.00
  1996,Jeep,Grand Cherokee,"MUST SELL!
  air, moon roof, loaded",4799.00
  2000,Toyota,Camry,,6700.0

- Example SQL statement:

  ```
  SELECT *
  FROM TABLE ( HDFS_Read('http://bigins.example.com/path/to/file.csv', '') )
  AS TX (YEAR INT, Make VARCHAR(10), Model VARCHAR(30),
  Description VARCHAR(40));
  ```

UDFs

- Result

| YEAR	MAKE	MODEL	DESCRIPTION
1997	Ford	E350	ac, abs, moon
1999	Chevy	Venture "Extended Edition"	(empty string)
1996	Jeep	Grand Cherokee	MUST SELL! air, moon roof, loaded
2000	Toyota	Camry	(null)

JAQL_SUBMIT

- JAQL_SUBMIT function enables invocation of IBM BigInsights Jaql from a DB2 application.
    - JAQL_SUBMIT(jaql-script, return-string, url, timeout-string)
- **jaql-script -** contains one or more Jaql queries, with or without parameter declarations.
- **return-string -** location for the output of the Jaql script.
- **url -** of a Jaql server that accepts requests from this DB2 client, executes those queries, and returns results.
- **timeout-string -** Maximum time (in seconds) to wait for results from the Jaql server.
- Result of the function is VARCHAR(512). Upon successful completion of the Jaql script, the function returns the return-string which is the location for the output of the Jaql script so that the Jaql_Submit UDF could be nested inside the HDFS_Read.
- If the Jaql script fails then the empty string will be returned.
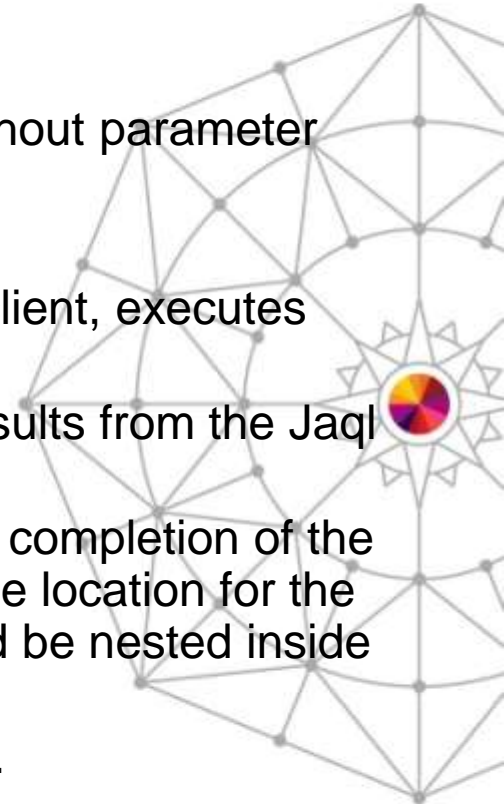
An example

```
Select *
From Table(hdfs_read(Jaql_submit('read(hdfs("receipts.dat"))
  -> transform { seller: $.Emisor.rfc,
                 amount: $.Conceptos }
  -> group by $seller =$.seller
     into {$seller, total:sum($.amount))
  -> write (del(location=''/tmp/test1.csv''))',
  'http://kea.svl.ibm.com:8080/data/controller/dfs/tmp/test1.csv',
  'http://kea.svl.ibm.com:8080', '1000'), '') )
  AS TX (seller  varchar(10),
         total   decfloat );
```
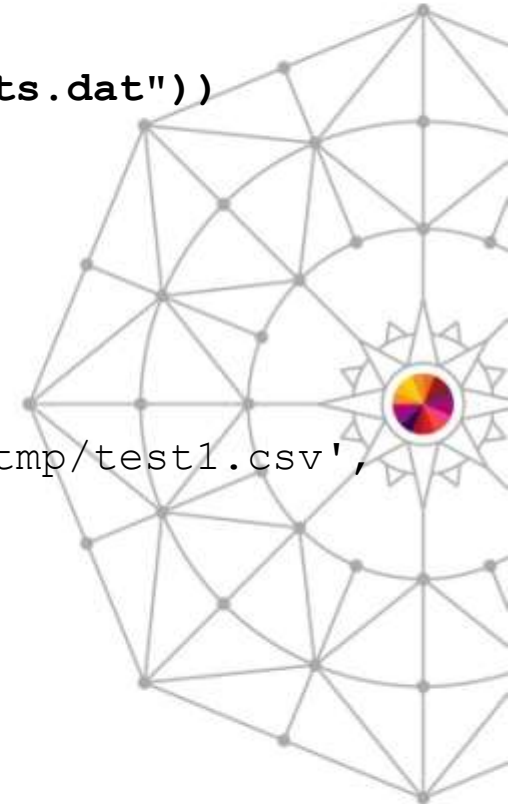
What's Next

- Future Phases
  - Use Emerging Big SQL Standard
    - DB2 for zOS will interact with HDFS Big SQL environments

  - Efficient bulk data transfer from DB2 to BigInsight

  - Efficient bulk data transfer from BigInsight to DB2 tables:
    - Jaql JDBC module to insert data into zDB2 using multiple Hadoop Reduce tasks
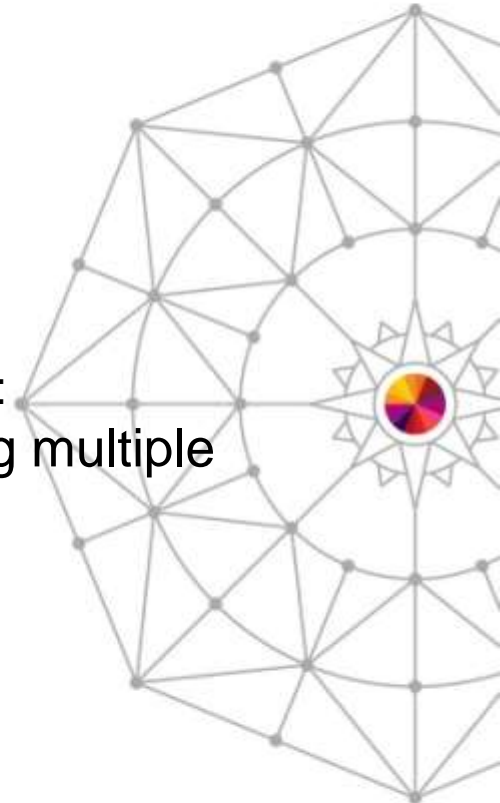
Section Summary

- Overview IBM BigData platform
- Use cases of integrating Bigdata with DB2 for zOS
- Functions to access bigInsight from DB2 V11
  - Generic Table Function
  - Big Data Job Submit udf based on BI REST API
  - HDFS_Read to read file from bigInsight based on BI REST API
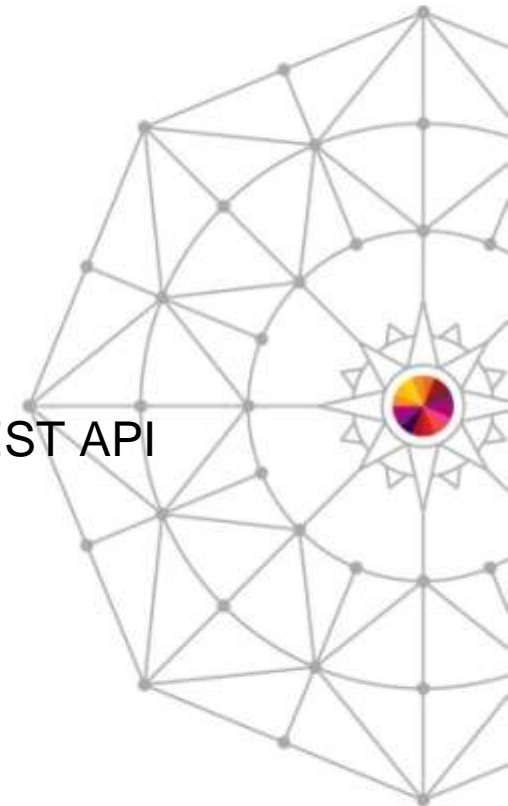
Agenda

- Big Data in an Information Driven economy
- Why start with System z
- DB2 for z/OS and InfoSphere BigInsights /Hadoop
- Summary / Call to action
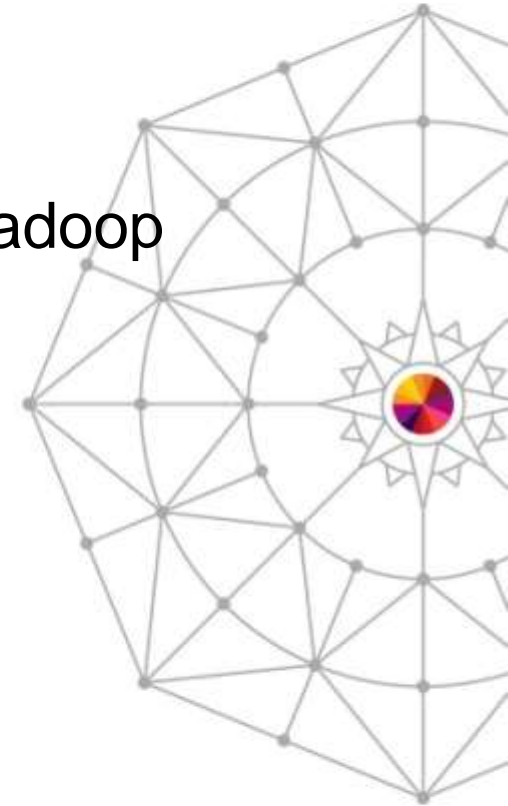
Slashing Time to Value / Gaining deeper insights
with IBM zEnterprise

Banco do Brasil purchases the **largest ever** DB2 Analytics Accelerator solution to drive customer insight from operational data. The 120-way system can hold 1.28 Petabytes of data. Queries that previously took *11 hours to run now complete in 26 seconds*, over 1500 times faster!

Banca Carige chooses System z to provide real time analytics as part of their Big Data client solution

BANCA CARIGE
Cassa di Risparmio di Genova e Imperia

Big Data at Banca Carige

Implementing a Mission Cr[...]

Big Data Exploration

Find, visualize, understand
all big data to improve
decision making

GPS & sensor information volumes exceeded the capabilities of the
existing system. It was redesigned as an enterprise mission critical
application using DB2 for z/OS and System z data sharing to now provide
the availability and scalability to meet the current and future requirements
for this solution.

JOHN DEERE

SHARE in Anaheim

Take Action Now!

Conclusion / Call to action:

- **For additional information including whitepapers and demos, please visit:**
  - IBM big data for z web site
  - Smarter Computing
  - Information Management System z

- **Education**
  - Free online education at bigdatauniversity.com
  - 145,000+ registered students

- **Further developments:**
  - Future webcast and announcements
  - World of DB2 for z/OS

- **Wanting to experiment on a big data integration project ? *Partner with IBM Silicon Valley Laboratory. (mtsimmon@us.ibm.com / jacopi@us.ibm.com)***

- **Develop your own big data strategy** –Contact your local IBM sales representative to get started.

Two Analyst Webcasts

"Big Data : Past, Present and Future - *An analyst's perspective*"

Register : http://ibm.co/1cnWZ32

"Big data and the enterprise – *A perspective from featured Gartner analyst Donald Feinberg*"

Register : http://bit.ly/GARTNERBIGDATA
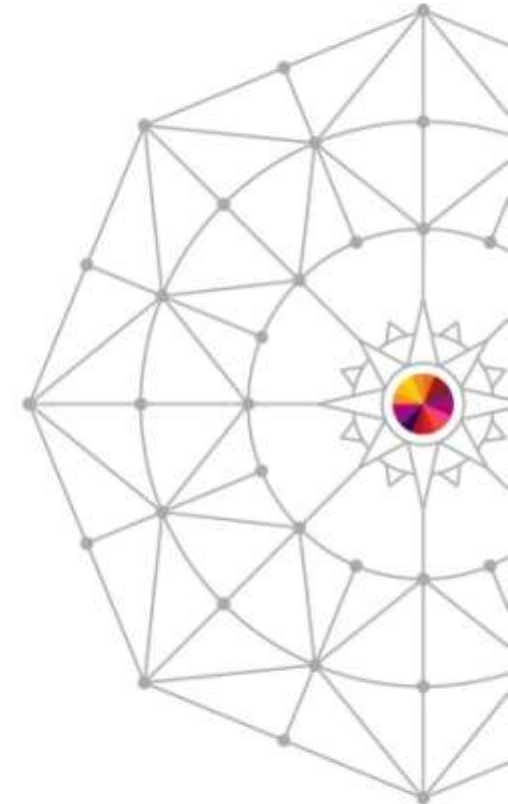
Useful URLs

- *ibm.com/bigdata/z*
- *ibm.com/smarterplanet*
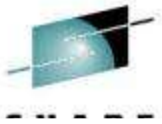- *ibm.com/software/os/systemz*
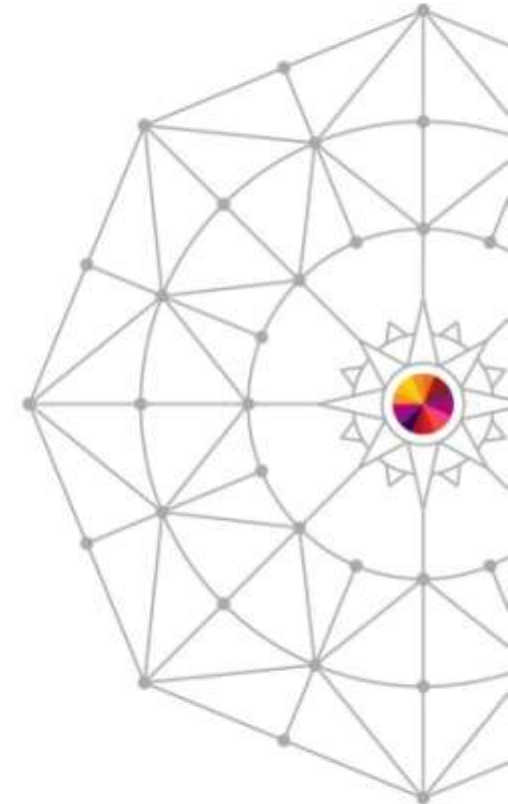
THINK
BIG

THINK

Z

Big Data Strategies with DB2 for z/OS
IDZ-3640

Mark Simmonds – WW Product Marketing and IT architect

Joachim Limburg – Director DB2 for z/OS

IBM Corporation

Thank You

Your feedback is important!

- Access the Conference Agenda Builder to complete your session surveys

  o Any web or mobile browser at
  http://iod13surveys.com/surveys.html

  o Any Agenda Builder kiosk onsite

SHARE in Anaheim