IBM Big Data Platform
The big data opportunity

Extracting insight from an immense volume, variety and velocity of data, in a timely and cost-effective manner.

**Variety:** Manage the complexity of multiple relational and non-relational data types and schemas

**Velocity:** Streaming data and large volume data movement

**Volume:** Scale from terabytes to zettabytes
What is “BIG DATA”?

All kinds of data

Large volumes

Valuable insight, but difficult to extract

Often extremely time sensitive
What makes big data technology different?

Jobs distributed across affordable hardware.
Manages and analyzes all kinds of data.
Analyzes data in native format.
Where do Enterprises get Stuck with Big Data?

Unwillingness to Change

Shortage Skilled Resources

One-Size-Fits-All

Lack Executive Support

Vague Goals

Processing Bottlenecks

The key to success in a Big Data world:
Organizational change to embrace exploratory analytics
IBM Big Data Platform Components

InfoSphere BigInsights
Hadoop-based low latency analytics for variety and volume

Hadoop

InfoSphere Information Server
High volume data integration and transformation

Information Integration

MPP Data Warehouse

IBM InfoSphere Warehouse
Large volume structured data analytics

IBM Netezza High Capacity Appliance
Queryable Archive Structured Data

IBM Netezza 1000
BI+Ad Hoc Analytics Structured Data

IBM Smart Analytics System
Operational Analytics on Structured Data

Stream Computing

InfoSphere Streams
Low Latency Analytics for streaming data
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
### IBM’s unique strengths in Big Data

| 1. **Big Data in Real-Time** | - Ingest, analyze and act on massive volumes of streaming data  
| | - Faster AND more cost-effective for specific use cases.  
| | (10x volume of data on the same hardware.) |
| 2. **Fit for purpose analytics** | - Analyzes a variety of data types, in their native format  
| | - For example: text, geospatial, time series, video, audio & more |
| 3. **Enterprise Class** | - High performance warehouse software and appliances  
| | - Ease of use with end users, admin and development UIs  
| | - Open source enhanced for reliability, performance and security |
| 4. **Integration** | - Integration into your IM architecture  
| | - Pre-integrated analytic applications |
InfoSphere BigInsights

INFORMATION INTEGRATION & GOVERNANCE FORUM 2012
Delivering Trusted Information for Smarter Business Decisions

© 2012 IBM Corporation
BigInsights delivers a comprehensive Hadoop-based big data solution.

**Volume**  
Terabytes, petabytes, even exabytes

**Variety**  
All kinds of data  
All kinds of analytics

**Velocity**  
Analyze data in...  
Hours instead of days  
Days instead of weeks

Store  
Analyze  
Explore

Traditional / Non-traditional data sources
BigInsights provides unique business value when:

- Data volumes cannot be cost effectively managed using existing technologies.
- Analyzing larger volumes of data can provide better results.
- Mining insights from non-relational data types.
- Exploring data to understand it’s potential value to the business.
What makes BigInsights special?

- **Scalable**
  - New nodes can be added on the fly.

- **Affordable**
  - Massively parallel computing on commodity servers

- **Flexible**
  - Hadoop is schema-less, and can absorb any type of data.

- **Fault Tolerant**
  - Through MapReduce software framework

- **Performance & reliability**
  - Adaptive MapReduce, Compression, BigIndex, Flexible Scheduler

- **Analytic Accelerators**

- **Productivity Accelerators**
  - Web-based UIs
  - Tools to leverage existing skills
  - End-user visualization

- **Enterprise Integration**
  - To extend & enrich your information supply chain.
InfoSphere BigInsights – Hadoop-Based Big Data Solution

Hadoop

A set of capabilities to cost-effectively analyze a wide variety and large volume of information to gain insights that were not previously possible. Ability to analyze data in its native format, without imposing a schema/structure, to enable fast ad-hoc analysis.

The big data platform builds on the Apache Hadoop project and incorporates latest technology from IBM research.

Use Cases

- **Analyze a Variety of Information**
  Novel analytics on a broad set of mixed information that could not be analyzed before

- **Analyze Extreme Volumes of Information**
  Cost-efficiently process and analyze petabytes of information

- **Discovery and Experimentation**
  Quick and easy sandbox to explore data and determine its value
Hadoop Explained

- Hadoop computation model
  - Data stored in a distributed file system spanning many inexpensive computers
  - Bring function to the data
  - Distribute application to the compute resources where the data is stored

- Scalable to thousands of nodes and petabytes of data

MapReduce Application

Distribute map tasks to cluster

Hadoop Data Nodes

1. Map Phase
   (break job into small parts)

2. Shuffle
   (transfer interim output for final processing)

3. Reduce Phase
   (boil all output down to a single result set)
InfoSphere BigInsights – A Full Hadoop Stack
How Text Analytics Works

Football World Cup 2010, the Dutch team distinguished themselves well, losing to the Spanish champions 1-0 in the Final. Early in the second half, Netherlands’ striker, Arjen Robben, had a breakaway, but the keeper for Spain, Iker Casillas made the save. Winger Andres Iniesta scored for Spain for the win.

<table>
<thead>
<tr>
<th>Name</th>
<th>Position</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arjen Robben</td>
<td>Striker</td>
<td>Netherlands</td>
</tr>
<tr>
<td>Iker Casillas</td>
<td>Keeper</td>
<td>Spain</td>
</tr>
<tr>
<td>Andres Iniesta</td>
<td>Winger</td>
<td>Spain</td>
</tr>
</tbody>
</table>
What is Text Analytics?

Text Analytics

Parsing Facts Deep understanding of text

Social media Call center records Logs SEC filings
InfoSphere Streams

INFORMATION INTEGRATION & GOVERNANCE FORUM 2012
Delivering Trusted Information for Smarter Business Decisions
InfoSphere Streams delivers Real Time Analytic Processing

A Platform to Run In-Motion Analytics on BIG Data

**Volume**
- Terabytes per second
- Petabytes per day

**Variety**
- All kinds of data
- All kinds of analytics

**Velocity**
- Insights in microseconds

Real time delivery

**Powerful Analytics**
- ICU Monitoring
- Environment Monitoring
- Algo Trading
- Cyber Security
- Government / Law enforcement
- Smart Grid
- Telco churn predict

**Traditional / Non-traditional data sources**

**Microsecond Latency**

**Millions of events per second**
Big Data in Real-Time with InfoSphere Streams
Streams provides unique business value when:

- It would be too expensive to store before analyzing.

- Data fusion across multiple, disparate streams brings advantage.

- True real-time data analysis can provide better business outcomes.

- Ability to run multiple analytic models or applications against the same data.
How it works:

Streams Processing Language (SPL)
built for Streaming apps:
- Reusable operators
- Rapid application development
- Continuous “pipeline” processing

Easy to extend:
- Built in adapters
- Custom developments with C++ and Java

Easy to manage:
- Automatic placement
- Extend apps incrementally w/out downtime
- Multi-user / multiple applications

Any type of data:
Can handle virtually any data type.

Compile groups of operators into single processes:
- Efficient use of cores
- Distributed execution
- Very fast data exchange
- Can be automatic or tuned
- Scaled with push of a button

Dynamic analysis:
- Programmatically change topology at runtime
- Create new subscriptions
- Create new port properties

Flexible & high performance transport:
- Ultra low latency
- High data rates
Accelerators to speed and simplify development

Telco

Finance

Smarter energy

Data mining

Public transportation

. . . with more on the way.

Over 100 samples applications

User Defined Toolkits

Standard Toolkit
>300 functions & operators
Stream Computing
A platform for managing information streams and applying big data analytics to data in motion. Analyze a wide variety of data in its native format, at a massive volume and scale (terabytes per second).

Use Cases

**Analyze Information in Motion**
Low-latency analysis of streaming information

**Analyze Extreme Volumes of Information in Motion**
Terabytes per second, petabytes per day.

**Analyze a Variety of Information**
Analyze a variety of data in its native format – streaming audio, video, spatial, among others
IBM InfoSphere Streams v2.0

**Agile Development Environment**
- Eclipse IDE
- Streams Live Graph
- Streams Debugger

**Distributed Runtime Environment**
- Clustered runtime for near-limitless capacity
- RHEL v5.3 and above
- x86 multicore hardware
- InfiniBand support

**Sophisticated Analytics with Toolkits & Adapters**
- Database
- Mining
- Financial
- Standard
- Internet
- Big Data (HDFS)
- Text
- User-defined toolkits
IBM Big Data Platform
What Could *You* do with a Big Data Platform?

**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
THINK
BIG