Big Data and Power BI

Darwin Schweitzer
Contact Info

Darwin Schweitzer

Email: darsch@microsoft.com
Github: datasnowman
Twitter: @DataSnowman
LinkedIn: Cortana Intelligence Solutions
Big data and analytics has huge potential

90% of business leaders expect big data to dramatically change how they do business, *having a landmark affect similar to the development of the Internet.*  
Source: Accenture, 2016

5X improvement in decision making by using advanced analytics  
Source: Bain & Co., 2016

6-8% productivity gains that result from investment in data analytics.  
"This is a higher rate of return than other recent technologies have yielded, surpassing even the computer investment cycle in the 1980s”  
#devdaysasia2017
#azure
#azureiot
#Microsoft

#cloudcomputing
#bigdata
@MSPowerBI
@MSAdvAnalytics
#iot
Time Series Insights using Twitter

Big Data is driving transformative changes

<table>
<thead>
<tr>
<th>Data characteristics</th>
<th>Traditional</th>
<th>Big Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data</td>
<td>Relational data with highly modeled schema</td>
<td>All data with schema agility</td>
</tr>
<tr>
<td>Costs</td>
<td>Specialized HW</td>
<td>Commodity HW</td>
</tr>
<tr>
<td>Culture</td>
<td>Operational reporting Focus on rear-view analysis</td>
<td>Experimentation leading to intelligent action With machine learning, graph, a/b testing</td>
</tr>
</tbody>
</table>
Big Data as a cornerstone of Cortana Intelligence
Microsoft Big Data

HDInsight

Data Lake Store

Data Lake Analytics

HDFS

YARN

.NET

Java

Scala

R

C#
Why Hadoop in the Cloud?

- No HW costs
- Unlimited scale
- Pay what you need
- Deployed in minutes

Microsoft
Something for all roles
Lower total cost of ownership

- No hardware
- Hadoop support included with Azure support
- Pay only for what you use
- Independently scale storage and compute
- No need to hire specialized operations team
- 63% lower total cost of ownership than on-premises*

*IDC study “The Business Value and TCO Advantage of Apache Hadoop in the Cloud with Microsoft Azure HDInsight”
Easy for business analysts with interactive reports over big data

- Interactive BI with big data
- Spark 2.0 integration
- Interactive Hive with LLAP - keeps data compressed running in-memory 25x faster
- ODBC driver to use Power BI or third party tools (Tableau, SAP, Qlik, etc.)
Easy for data scientists with familiar R language

R Server for HDInsight
- Largest portable R parallel analytics library
- Terabyte-scale machine learning—1,000x larger than in open source R
- Up to 100x faster performance using Spark and optimized vector/math libraries
- Enterprise-grade security and support

*Applies to HDInsight only
Easy notebook experience for data engineers

- Most popular notebooks, Jupyter and Zeppelin out-of-the-box
- Combine code, statistical equations and visualizations
- Worked w/ Jupyter community to enhance kernel to allow Spark execution through REST endpoint
Debug and Optimize your Big Data programs with ease

- Deep integration with IDEs for developer productivity: Visual Studio, Eclipse, & IntelliJ
- Integrated with Hive, Pig, Storm, and Spark
- Visually see execution of Hive jobs ran by the Tez execution engine
- Full Intellisense
Easy for administrators to spin up quickly

- Deploy big data projects in minutes
- No hardware to install, tune, configure or deploy
- No infrastructure or software to manage
- Scale to tens to thousands of machines instantly
Recognized by top analysts

Forrester Wave for Big Data Hadoop Cloud

- Named industry leader by Forrester with the most comprehensive, scalable, and integrated platforms*
- Recognized for its cloud-first strategy that is paying off*

*The Forrester Wave™: Big Data Hadoop Cloud Solutions, Q2 2016.
ISV Integration

Integration with leading productivity applications
Spin up Hadoop and Spark clusters pre-integrated and pre-tuned with ISV applications out-of-the-box
Runs on the HDInsight clusters; does not require separate VMs
Fast and easy way to spin up applications
Azure Data Lake Store

A No limits Data Lake that powers Big Data Analytics

- Petabyte size files and Trillions of objects
- Scalable throughput for massively parallel analytics
- HDFS for the cloud
- Always encrypted, role-based security & auditing
- Enterprise-grade support
Azure Data Lake Analytics

A No limits Analytics Job Service to power intelligent action

Start in seconds, scale instantly, pay per job

Develop massively parallel programs with simplicity

Debug and optimize your big data programs with ease

Virtualize your analytics

Enterprise-grade security, auditing and support
Petabyte size files and Trillions of objects

- Store data in its native format
- PB sized files, \(200\times\) larger than anyone else
- Scalable throughput for massively parallel analytics
- No need to redesign application or reparation data at higher scale
Any type of analytics

- Batch, interactive, streaming, machine learning
- Allows for exploratory analytics over data
- Analyze with Hadoop and Microsoft solutions

Diagram:
- Analytics
  - U-SQL
- HDInsight
  - Hive
  - Spark
  - Storm
  - R Server
  - Kafka
- YARN
- HDFS
- Store
- Cortana Intelligence Suite
Start in seconds, Scale instantly, Pay per job

- Process big data jobs in 30 seconds
- No infrastructure to worry about (no servers, no VMs, no clusters)
- Instantly scale analytic units up or down (processing power)
- Architected for cloud scale and performance
- Frees you up to focus only on your business logic
Develop massively parallel programs with simplicity

- **U-SQL**: a simple and powerful language that’s familiar and easily extensible
- Unifies the declarative nature of SQL with expressive power of C#
- Leverage existing libraries in .NET languages, R and Python
- Massively parallelize code on diverse workloads (ETL, ML, image tagging, facial detection)

```sql
@t = EXTRACT date string , time string , author string , tweet string
    FROM "/Input/MyTwitterHistory.csv"
    USING Extractors.Csv();
@res = SELECT author AS author , COUNT(*) AS tweetcount
    FROM @t
    GROUP BY author;
OUTPUT @res TO "/Output/MyTwitterAnalysis.csv"
ORDER BY tweetcount DESC
    USING Outputters.Csv();
```
Azure SQL Data Warehouse

A relational **data warehouse-as-a-service**, fully managed by Microsoft. Industries first **elastic** cloud data warehouse with proven SQL Server capabilities. Support your **smallest to your largest** data storage needs.

- **Elastic scale & performance**
  - Scales to petabytes of data
  - Massively Parallel Processing
  - Instant-on compute scales in seconds
  - Query Relational / Non-Relational

- **Get started in minutes**
- **Integrated with Azure ML, PowerBI & ADF**

- **Market Leading Price & Performance**
  - Simple billing compute & storage
  - Pay for what you need, when you need it with dynamic pause
A fully managed Platform-as-a-Service

- Azure cloud data warehouse service
- Elastic scale
- Separate storage and compute
- Use existing tools and skills
- Deploy and use in minutes!
Remove the complexity of big data

T-SQL over Hadoop

Leverage existing T-SQL skills and BI tools to gain insights from different data stores

Key Benefits

• Reduce development costs performing query in structured and non-structured data using T-SQL
• Combine SQL with Hadoop enabling database for modern applications
• Simplified architecture with built-in hybrid integration
• Leverage existing T-SQL skills and BI tools to gain insights from different data stores

PolyBase

<table>
<thead>
<tr>
<th>Name</th>
<th>DOB</th>
<th>State</th>
</tr>
</thead>
<tbody>
<tr>
<td>Denny Usher</td>
<td>11/13/58</td>
<td>WA</td>
</tr>
<tr>
<td>Gina Burch</td>
<td>04/29/76</td>
<td>WA</td>
</tr>
</tbody>
</table>

SQL Server  Hadoop

Quote:
$658.39
Data Warehousing and Modern BI on Azure

Demo

Capture
Data Sources
- Ingest

Curate
- Sanitize
- Warehousing
- Caching
- Visualize

Consume
- Power BI

Business apps
- Writes to Transactional store

OLTP SQL Database
- Sync source tables at hourly interval to partitioned store in Blob

Azure Blob storage
- Generate Facts and Dimensions From Source Tables in Blob
- SQL Data Warehouse (Data Mart)
- Insert incrementals via Polybase

HDInsight Hive

Data Factory: Move data, orchestrate, schedule and monitor

Azure Analysis Service
- Scheduled Cache Refresh

Jupyter Notebook

Event Producers ➔ Collection ➔ Stream ingestion ➔ Transformation & analytics ➔ Stream ingestion ➔ Presentation and Action

- Applications
- Devices
- Sensors
- Web and Social

Cloud gateways (web APIs) ➔ Kafka for HDInsight ➔ HDInsight

- Spark for HDInsight
- Storm for HDInsight

Kafka for HDInsight ➔ Stream processing ➔ Storage Adapters

Live Dashboards ➔ Web/thick client dashboards ➔ Search and query ➔ Data Analytics ➔ Event hub

Data Lake Store

MongoDB SQL

Solr Azure

Search

HDInsight

HBase for HDInsight

Microsoft

Power BI: experience your data
Any data, any way, anywhere

Out-of-the-box SaaS content packs
Real time dashboards & interactive reports
Natural Language query
Custom visualizations
Native Office 365 integration

Cloud data
On-premises data

Power BI

Web
Mobile
Excel
Embedded
Cortana
∞ “…….”
Worldwide reach

18,000 Cities globally access the service daily
235 Countries & territories
43 Languages
Global coverage 19 datacenters worldwide
National Clouds
US Government
China
Germany

Shaded areas show countries with Power BI usage. The darker the color the greater the number of users.
Microsoft Power BI

MAY 2017
Editor’s choice

PROS
Very powerful tool with loads of data source connectors.
Very easy to use. Superb data visualization capabilities.

CONS
Divides data prep tools between desktop and web versions. Refresh cycle is only daily on free version.

BOTTOM LINE
A powerful and highly intuitive self-service business intelligence (BI) tool that’s cheap, has a short learning curve, and can be extended in all kinds of directions depending on how many other Microsoft platforms you’re willing to use.
February 2017

A Leader in the Magic Quadrant for BI & Analytics*

*Gartner “Magic Quadrant for Business Intelligence and Analytics Platforms,” by Rita L. Sailam, Cindi Howson, Carlie J. Idoine, Thomas W. Oestreicher, James Laurence Richardson, and Joao Tapadinas February 16, 2017

This graphic was published by Gartner, Inc. as part of a larger research document and should be evaluated in the context of the entire document. The Gartner document is available upon request from Microsoft. Gartner does not endorse any vendor, product or service depicted in its research publications, and does not advise technology users to select only those vendors with the highest ratings or other designation. Gartner research publications consist of the opinions of Gartner’s research organization and should not be construed as statements of fact. Gartner disclaims all warranties, expressed or implied, with respect to this research, including any warranties of merchantability or fitness for a particular purpose.
Power BI Premium

Sharing without boundaries

Unrestricted scale and performance

Flexibility to deploy in the cloud or on-premises

Dedicated resources in the cloud
Cortana Intelligence Solutions
Improve business processes and increase competitiveness

**Intelligent**
Interactive dashboards that help highlight actionable insights from data
Proven solution architectures that enable transformation of data into insights

**Flexible**
Deploy and operationalize a complex big data architecture with ease.
Scale and customize the solution to your needs

**Trusted**
Qualified partners to help with proof-of-concepts and operationalization
Rich technical guides to build an enterprise-grade solution
Cortana Intelligence Solutions

- Solution can be discovered from Microsoft.com
- Solutions and Tutorials from the Cortana Intelligence Gallery
  https://gallery.cortanaintelligence.com/solutions
- From aka.ms short urls https://aka.ms/cisolutions and https://aka.ms/citutorials
- AppSource
Solutions Implementation Partners

Implementation Partners we are training on these solutions
Thank You!

Questions?