

Big Data Technologies Compared ▶ June 2014

Agenda

- What is Big Data
- Big Data Technology Comparison Summary
- Other Big Data Technologies
- Questions

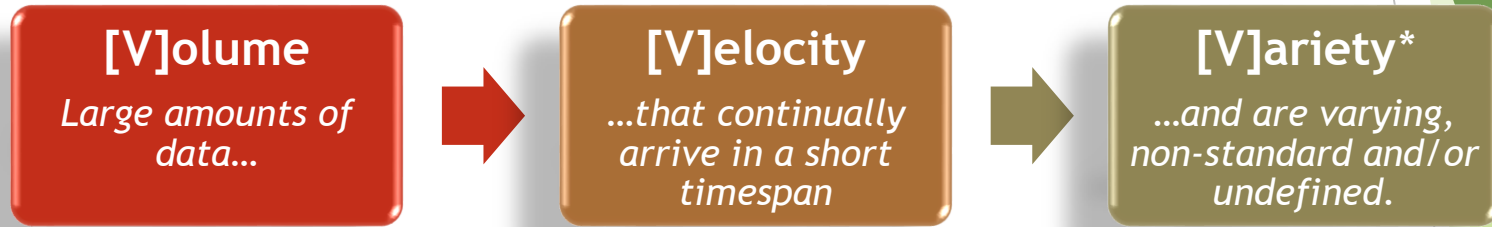
What is Big Data... *by Example*



- The **SKA Telescope** is a new development in South Africa & Australia comes online in **2020**
 - ▶ **Square Kilometre Array**
- “The SKA will generate enough raw data to fill 15 million 64GB iPods every day”
 - ▶ **913 PB per day** or **10.8 TB per sec** of RAW Unprocessed Data
 - ▶ *Post Processing generates a rather modest 10 GB per sec or 844 TB per day*
- But how do you process this **Super Data**? Naturally with a **Super Computer**!
- “The SKA will require very high performance supercomputers capable of 100 petaflops per second processing power. This is about 50 times more powerful than the most powerful supercomputer in 2010 and equivalent to the processing power of about 100 million PCs”

What is Big Data... *by Definition*

- Marketing term to describe **large, complex & varying** data sets that are **difficult to process**
- You tend to have a Big Data “problem” if your data sets have **these 3 properties...**

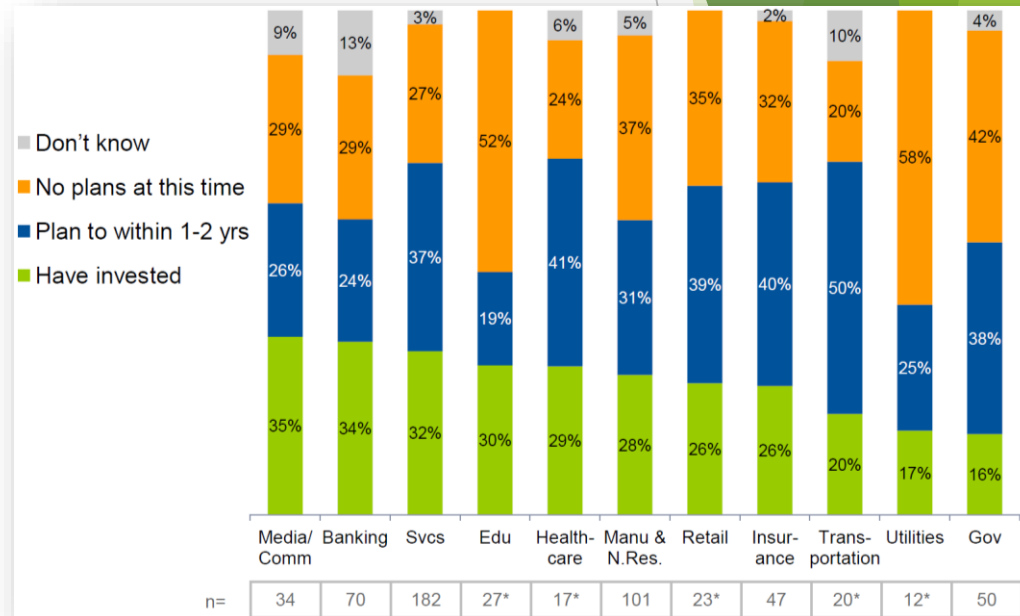


- To solve the “problem” you need specialist technologies
- Business want to solve the “problem” because *they believe* it offers competitive advantage
- **Traditional Database** is by the “problem” definition* **not** a Big Data technology or solution
- *However* traditional RDBMS vendors **extend** their products to **include** Big Data capabilities

What is Big Data... *by Market Size / Growth*

Growing 6 times faster than the overall IT market

- Gartner predicts the **big data** market will reach **\$16.1 billion in 2014** (incl. servers/storage)
- The **big data** professional services market will exceed **\$4.5 billion in 2014**
- The adoption of analytics-as-a-service will **accelerate ...**
(ie ready-made analytics in the cloud)
- The number of vendors providing **big data** services will **triple over the next three years**
- Different industries will leverage **Big Data** at different rates



Agenda

- What is Big Data
- Big Data Technology Comparison Summary
- Other Big Data Technologies
- Questions

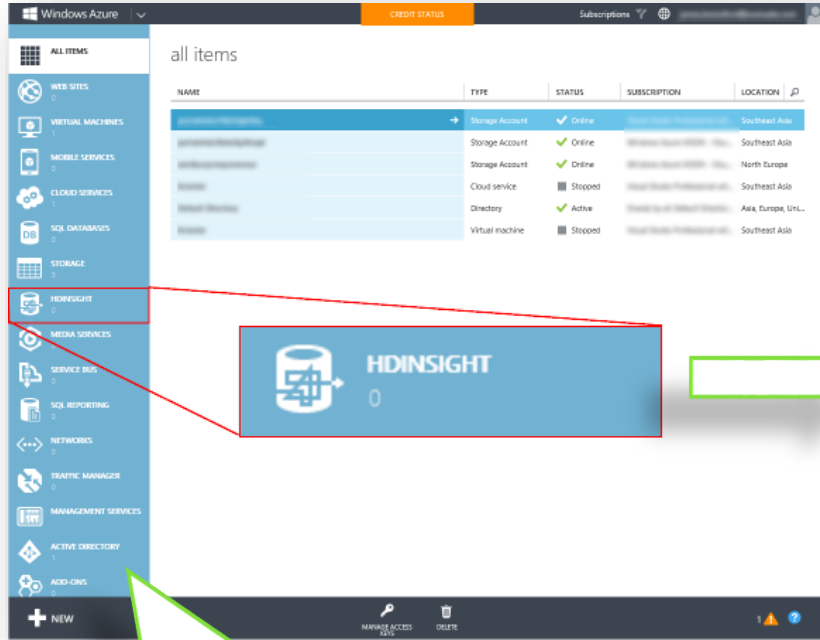
Big Data - Some Technology Comparisons - Summary

Factor // Technology	Microsoft HDInsight (Cloud based version of Hadoop)	Splunk	Microsoft APS (aka Microsoft SQL Server PDW)	EMC Greenplum	SAP Hana
Sweet Spot	Unstructured and Semi-structured data loading, storage, and query processing	Unstructured and Semi-structured data loading, storage, search, reporting and analytics applications	Structured tabular data warehousing applications and OLAP analytics	Structured tabular data warehousing applications and OLAP analytics	SAP Structured tabular OLTP applications
Who owns it	Microsoft	Splunk	Microsoft	EMC	SAP
Schema on...	On Read	On Read	On Write	On Write	On Write
RDBMS OOTB	No	No	Yes (native MPP)	Yes (native MPP)	Yes (in memory)
Hadoop OOTB	Yes (native MPP)	No (Hadoop Extn)	No (Hadoop Extn)	No (Hadoop Extn)	No (Hadoop Extn)
SQL Language	No (HIVE / Map Reduce)	No (SPL)	Yes (ANSI SQL)	Yes (ANSI SQL)	Yes (ANSI SQL)
Reports OOTB	No (MS & 3 rd party)	Yes & 3 rd party	No (MS & 3 rd party)	No (GP & 3 rd party)	No (BO & 3 rd party)
Format / Footprint	As-a-Service	Software & As-a-Service	Appliance	Software & Appliance	Appliance

Agenda

- What is Big Data
- Big Data Technology Comparison Summary
- Big Data Technology in Detail
- Other Big Data Technologies
- Questions

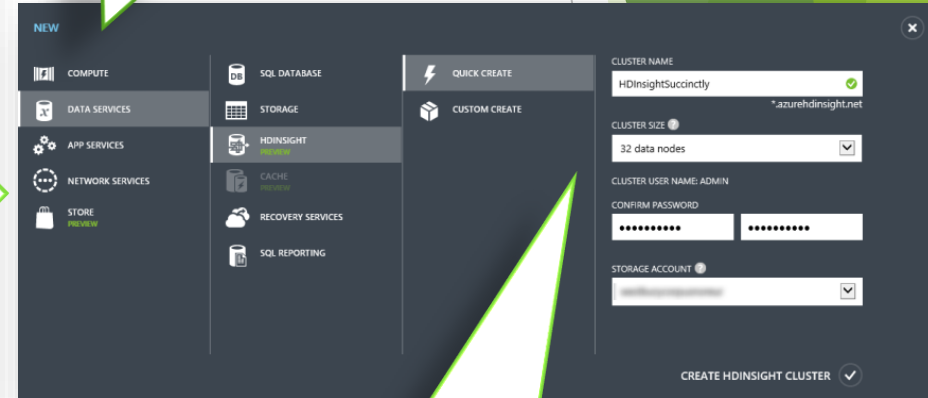
Microsoft HDInsight (Hadoop)



HDInsight is part of the range of Azure services

\$3,000/month for 10TB in 9 nodes

Can provision a 9 node cluster in less than 15 minutes



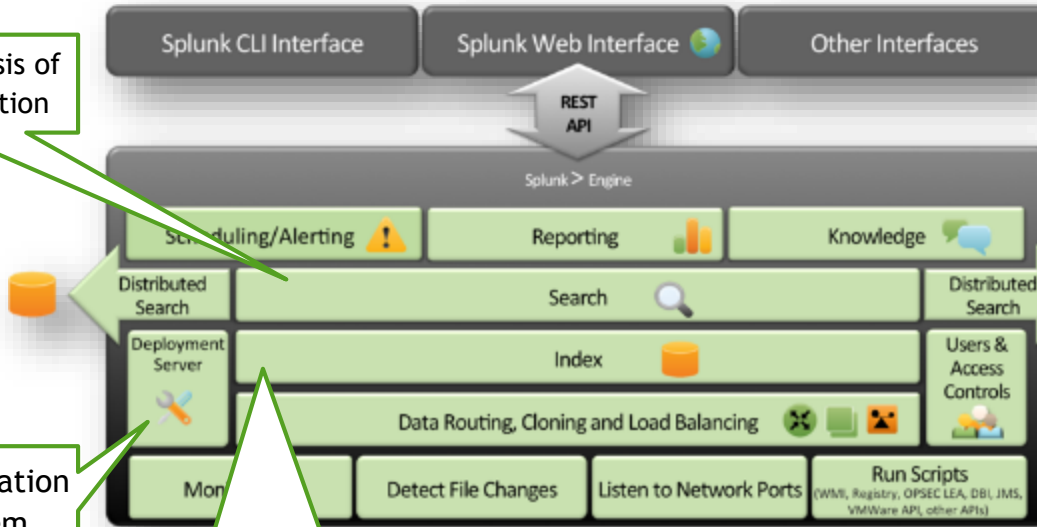
Cluster sizes from 4 to 32 nodes

Splunk

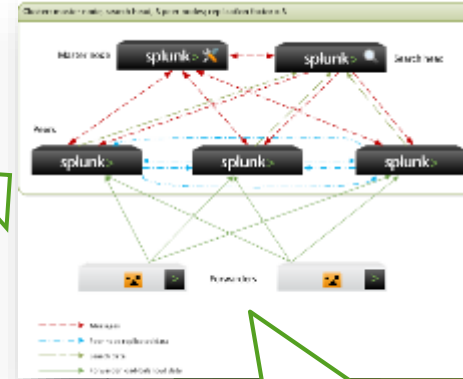
Search is the basis of all data interaction

Control configuration across a system (single & distributed)

Index data, and host primary data store, manage data aging (hot, warm, cold, frozen) based on age



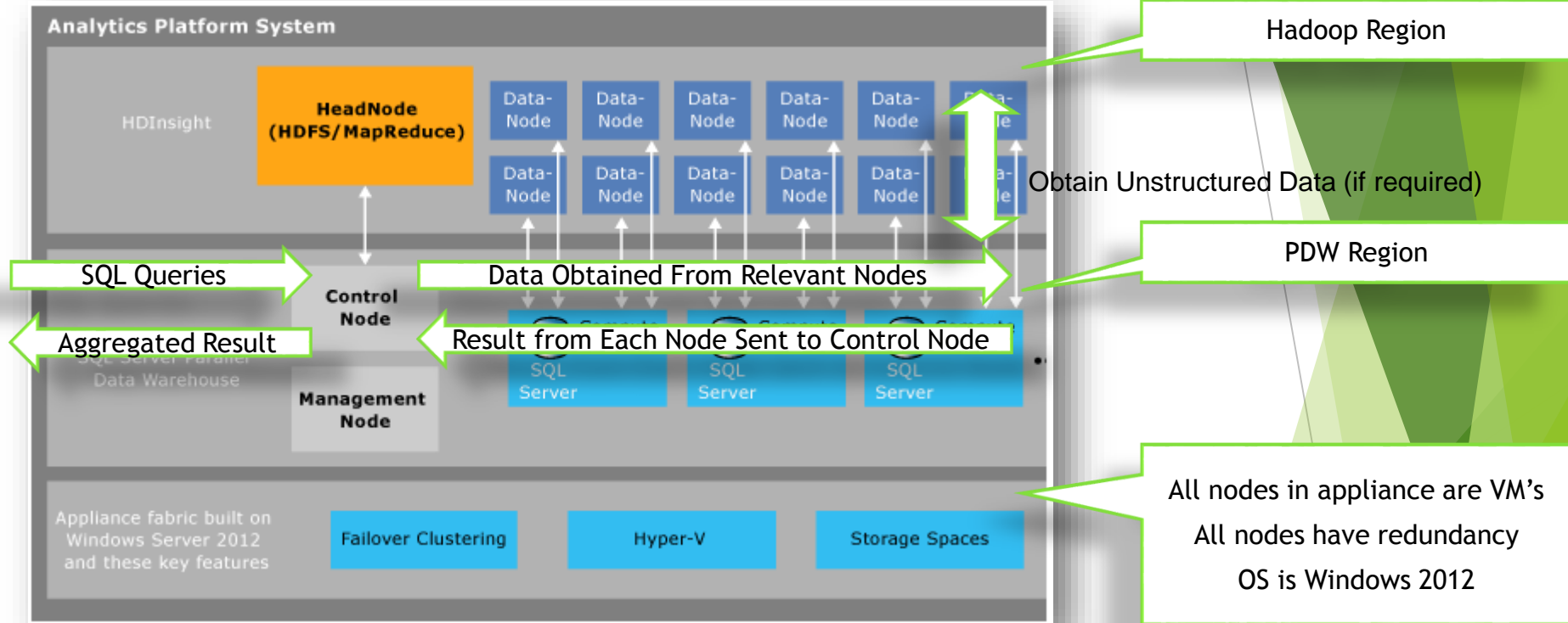
Cluster



Cluster architecture replicates copies of data for redundancy

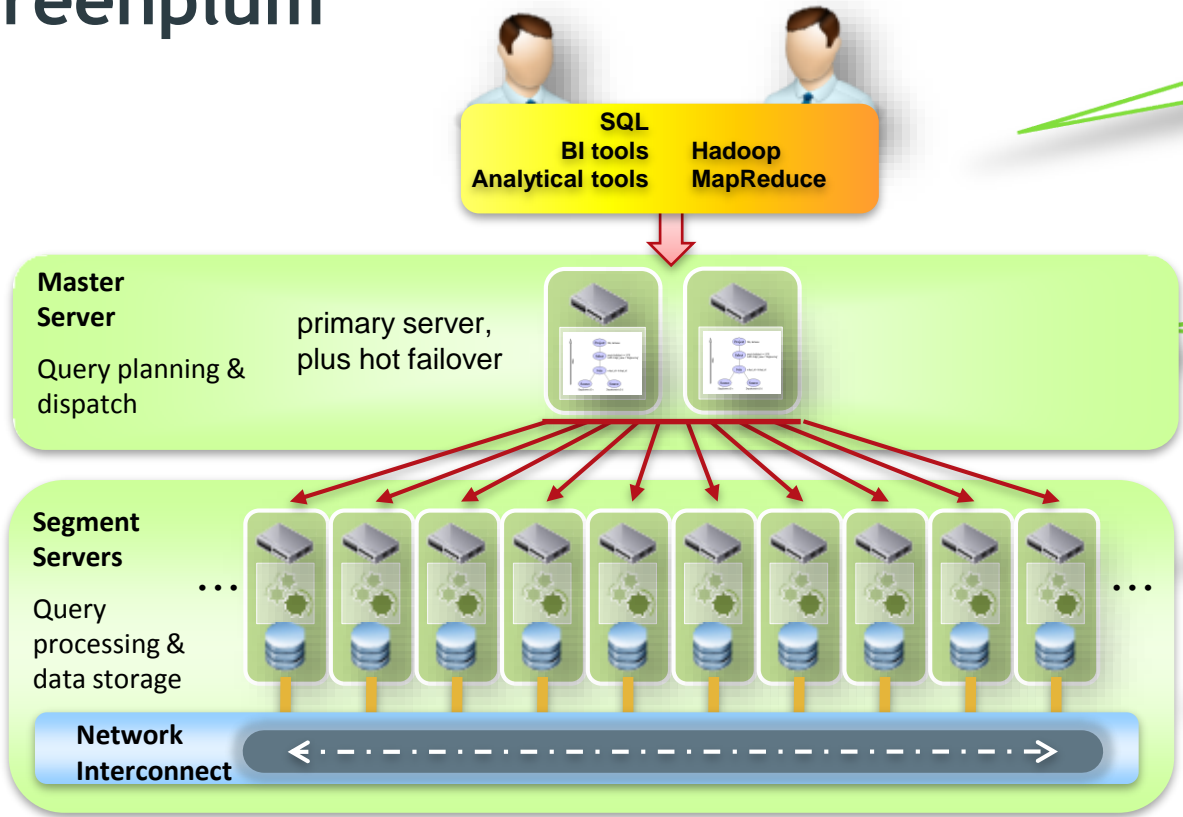
Distributed model for index, data store and search

Microsoft APS (PDW)



EMC Greenplum

Structured Analytics Unstructured Analytics



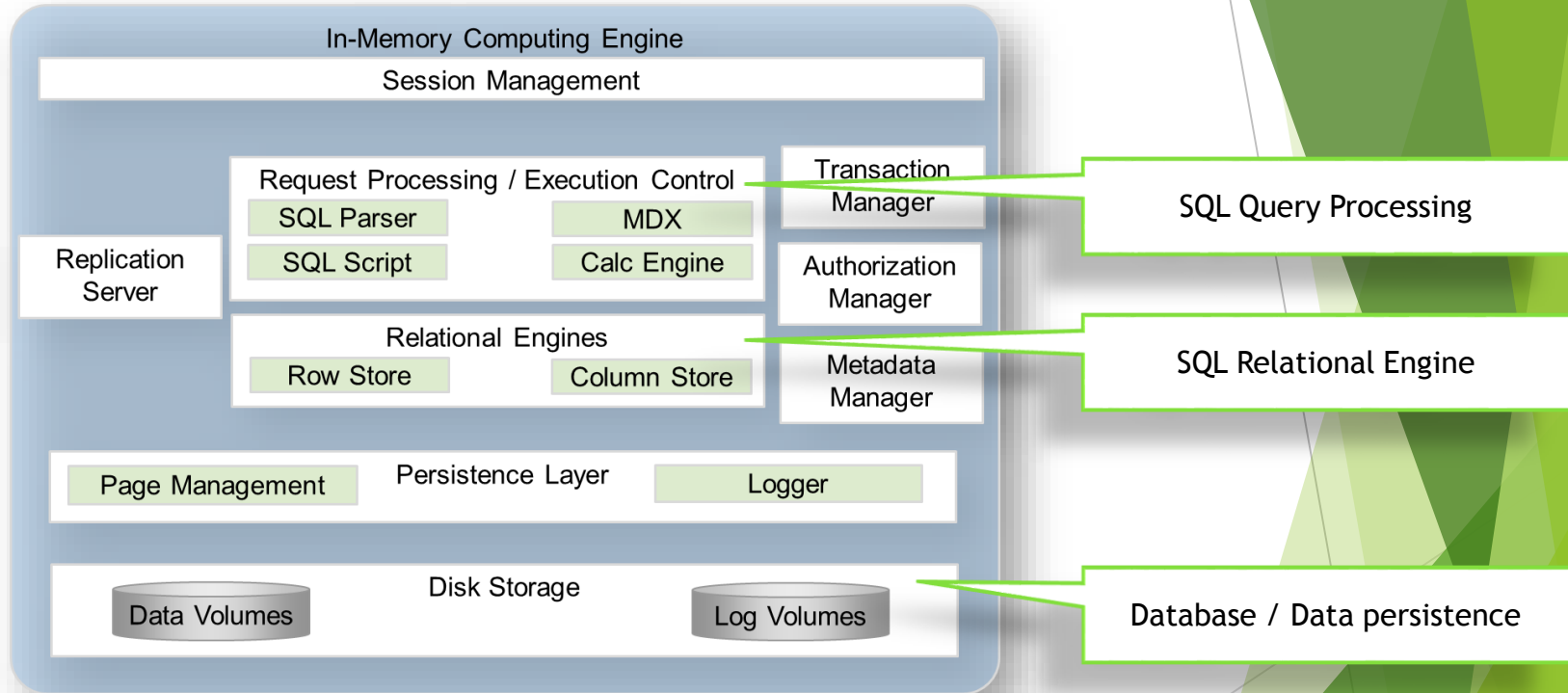
Standard Business Intelligence and Analytical tools

Clients see a single Postgres database

Queries distributed across all available resources

Shared Nothing, Massively Parallel Processing means no bottlenecks and linear scalability.

SAP Hana



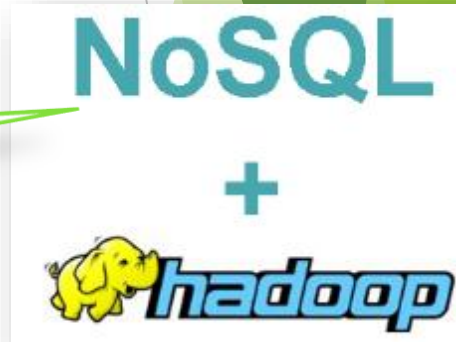
Agenda

- What is Big Data
- Big Data Technology Comparison Summary
- Other Big Data Technologies
- Questions

Other Technologies under the “Big Data” Banner You May Hear About

- Everyone who is anyone will probably claim to have a Big Data technology / solution
- Traditional Database Technologies (extended) → Tabular & Relational based technologies
 - Teradata, IBM DB2, Oracle Exadata, SAP/Sybase IQ, Netezza, ...
 - + *many more*
- NoSQL Technologies → Column, Key-Value Pair, Object and Graph based technologies
 - Casandra, MongoDB, Hbase, Allegro, BigTable, ...
 - + *many more*
- Hadoop Technologies
 - Cloudera, Hortonworks, Amazon, Google, Xplenty, ...
 - + *many more*

Not Only SQL



Agenda

- What is Big Data
- Big Data Technology Comparison Summary
- Other Big Data Technologies
- Questions



Questions...?

Appendix - References

- Big Data on Wiki
 - http://en.wikipedia.org/wiki/Big_data
- Gartner Data Warehousing Magic Quadrant
 - <http://www.gartner.com/technology/reprints.do?id=1-1RP452A&ct=140310&st=sb>
- The SKA Telescope
 - <http://www.skatelescope.org/the-technology/>
- Splunk on Wiki
 - <http://en.wikipedia.org/wiki/Splunk>
- NoSQL on Wiki
 - <http://en.wikipedia.org/wiki/NoSQL>
- The Big Data Marketplace
 - <http://www.forbes.com/sites/gilpress/2013/12/12/16-1-billion-big-data-market-2014-predictions-from-idc-and-iaa/>