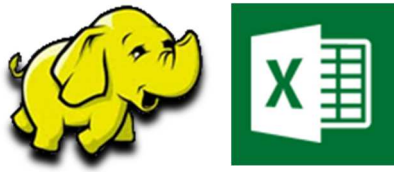
Berlin SQL Server User Group | 10 February, 2015



Big Data Analytics with Excel

Bitwise Solutions

people + technology + **training** = success

Presenter Introduction

Peter Myers

BI Expert – Bitwise Solutions

BBus, SQL Server MCSE, MCT, SQL Server MVP

Experienced in designing, developing and maintaining Microsoft database and application solutions, since 1997

Focuses on education and mentoring

Based in Melbourne, Australia

peter.myers@bitwisesolutions.com.au

http://www.linkedin.com/in/peterjsmyers

Session Outline

Introducing:

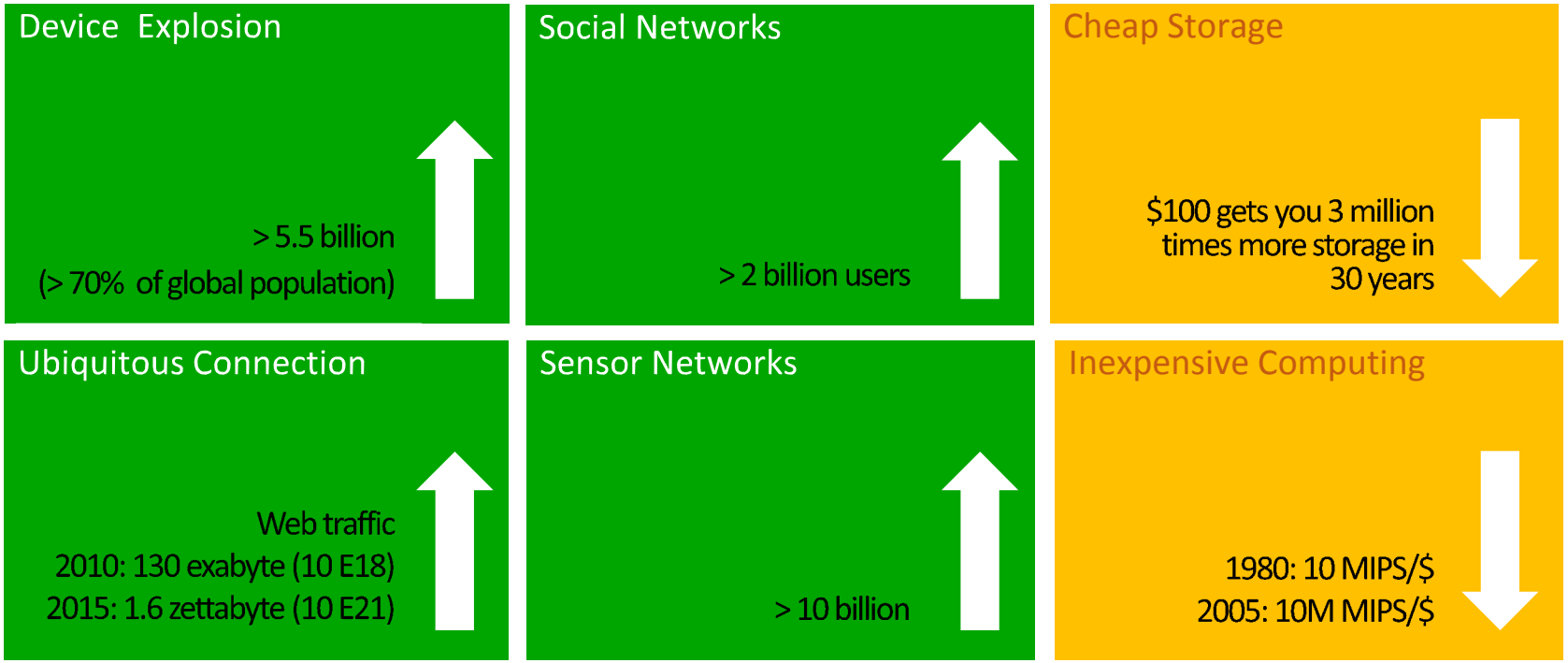
- Big Data
- Hadoop
- Azure HDInsight
- Power BI for Office 365

Big Data Modeling with Power Pivot:

- Benefits
- Considerations

Resources

Introducing Big Data



Introducing Big Data

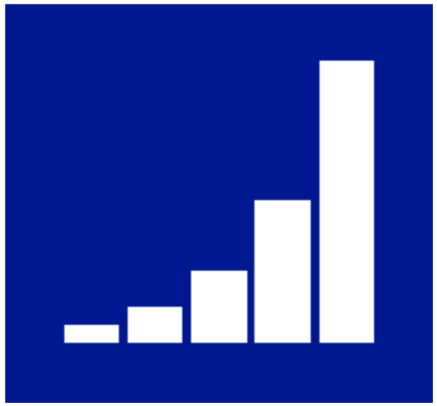
(Continued)

"Big data is a collection of data sets so large and complex that it becomes awkward to work with using on-hand database management tools. Difficulties include capture, storage, search, sharing, analysis, and visualization."

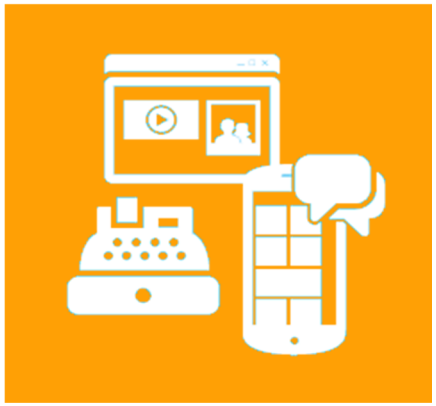
– Wikipedia

Introducing Big Data
(Continued)

Big data solutions deal with the complexities of:



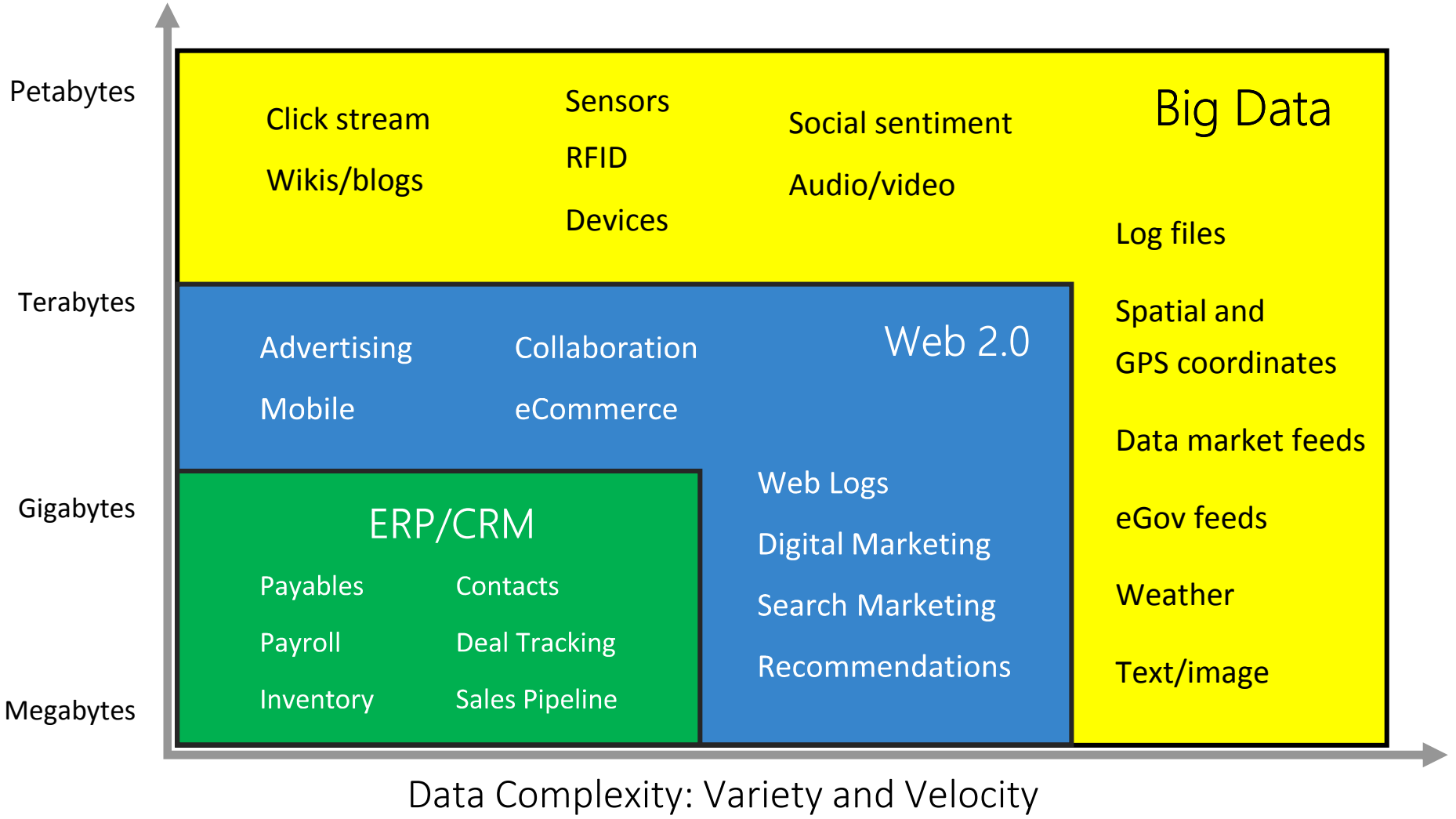
VOLUME
(Size)



VARIETY
(Structure)

VELOCITY
(Speed)

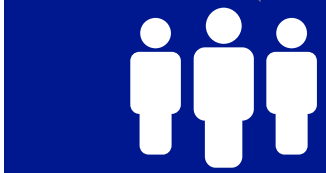
Introducing Big Data
(Continued)



Introducing Big Data

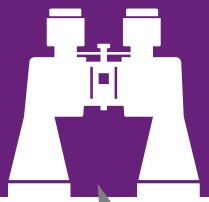
Responding to New Questions

What's the social sentiment of my product?



Social Analytics

Live Data Feed



How do I optimize my services based on patterns of weather, traffic, etc.?

How do I better predict future outcomes?

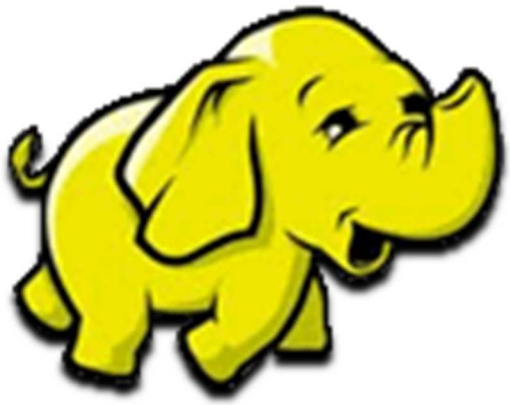
Advanced Analytics

Introducing Hadoop

Apache Hadoop is for big data

It is a set of open source projects that transform commodity hardware into a service that can:

- Store petabytes of data reliably
- Allow huge distributed computations



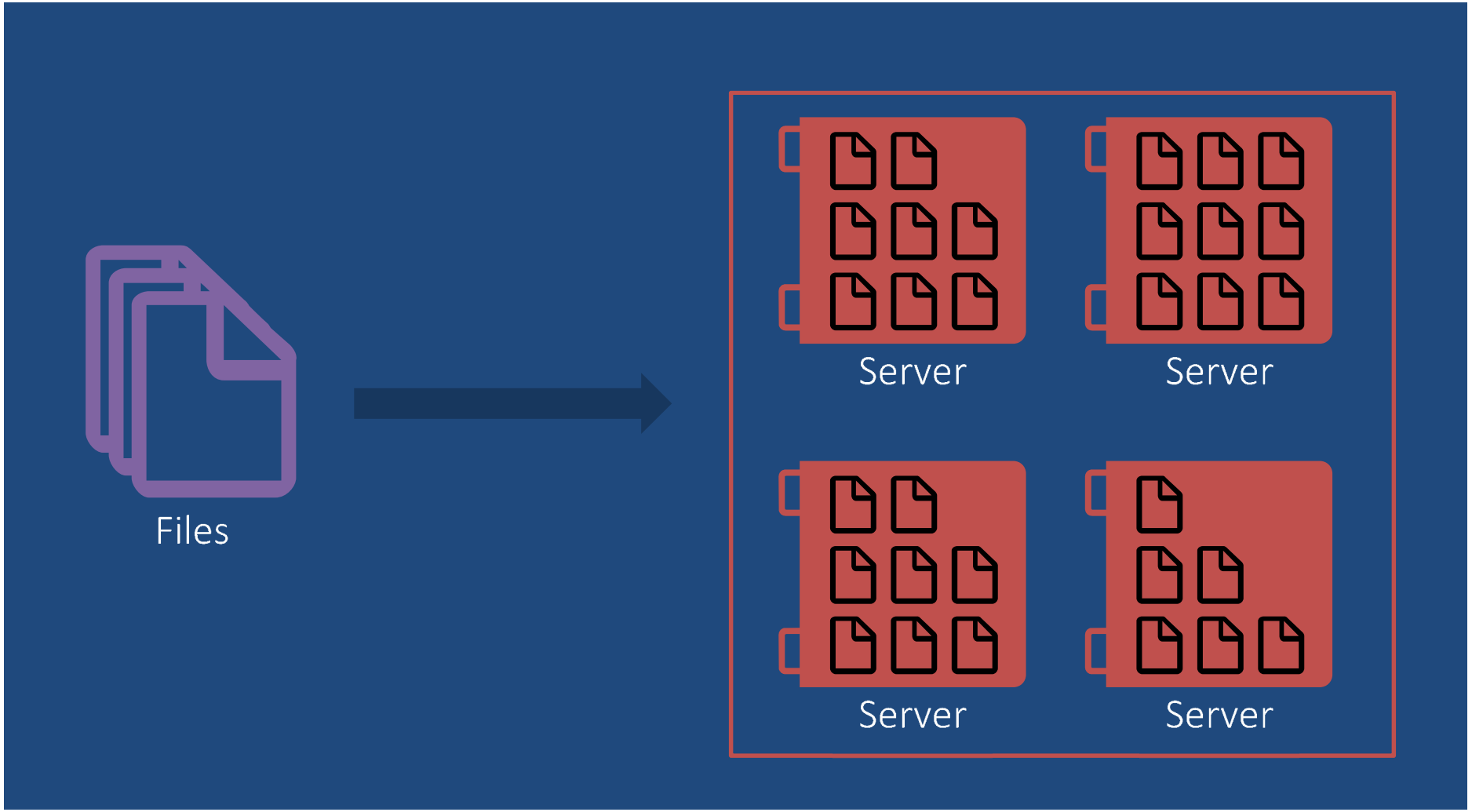
Introducing Hadoop
(Continued)

Key attributes:

- Open source
- Highly scalable
- Runs on commodity hardware
- Redundant and reliable (no data loss)
- Batch processing centric – using a “Map-Reduce” processing paradigm

Introducing Hadoop

How it Works: 1 – Data Storage

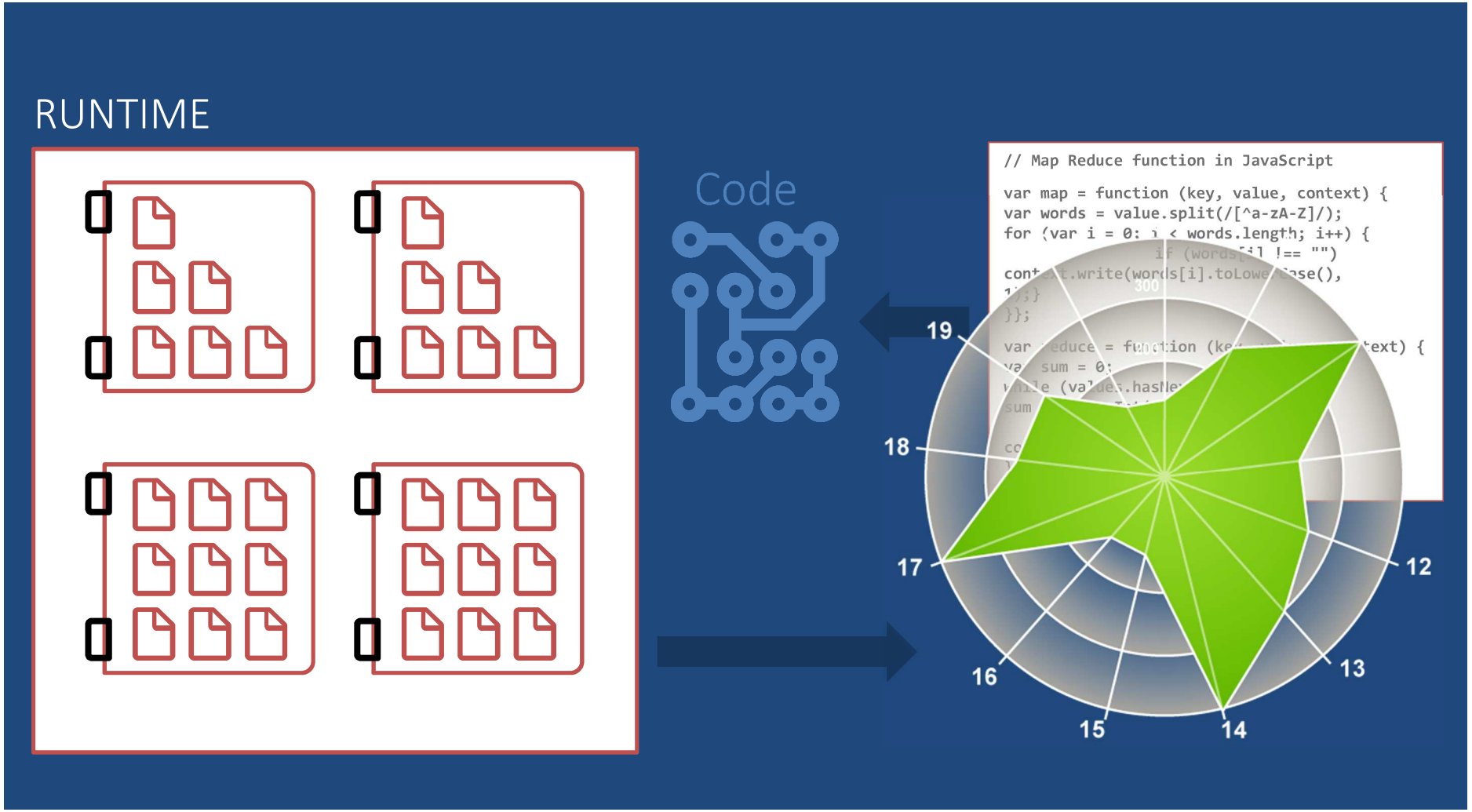


Demonstration

1. Storing Web Log Data

Introducing Hadoop

How it Works: 2 – Take the Processing to the Data



Introducing Hadoop

Comparison to Traditional RDBMS

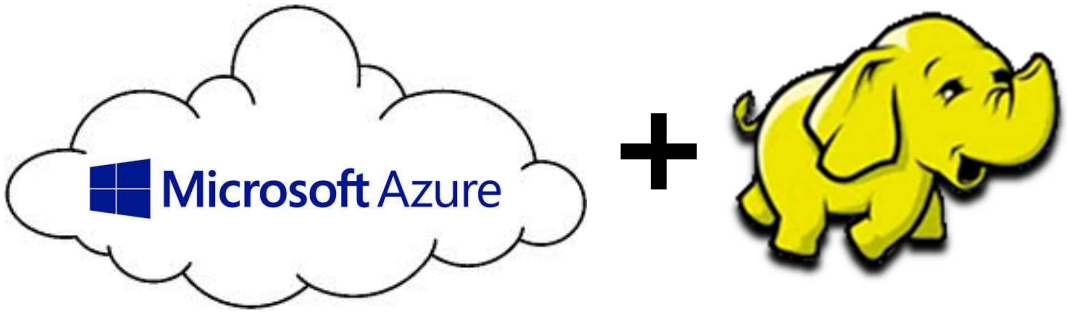
	TRADITIONAL RDBMS	HADOOP
Data Size	Gigabytes (Terabytes)	Petabytes (even Exabytes)
Access	Interactive and Batch	Batch
Updates	Read / Write many times	Write once, Read many times
Structure	Static Schema	Dynamic Schema
Integrity	High (ACID)	Low
Scaling	Nonlinear	Linear
DBA Ratio	1:40	1:3000

Reference: Tom White's Hadoop: The Definitive Guide

Introducing Azure HDInsight

Azure HDInsight is Microsoft's Hadoop-based service that enables big data solutions in the cloud

Empowers organizations with new insights on previously untouched unstructured data, while connecting to the most widely used BI tools on the planet



Introducing Azure HDInsight
(Continued)

Key attributes:

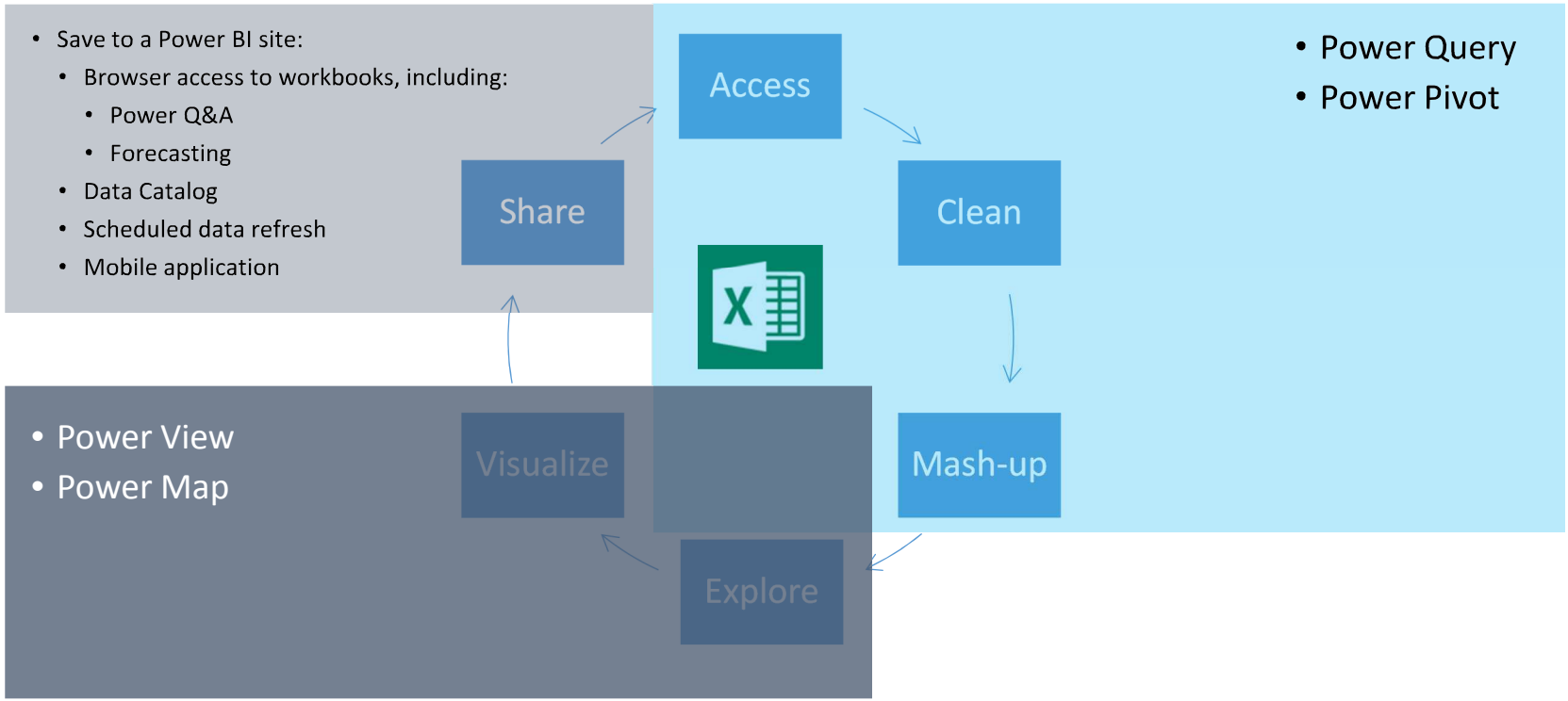
- 100% Apache Hadoop solution in the cloud
- Built on Hortonworks Data Platform (HDP)
- Deployment agility
- Develop in .NET and Java
- Can be automated with PowerShell and Command Line
- Can deliver insights through Excel

Demonstrations

1. Word count with **Pig** (“Hello World” for Hadoop)
2. Web Log count with **Hive**

Power BI for Office 365

Excel: Complete and Powerful Self-Service Tool



Demonstrations

1. Using Power Query to query the Pig output
2. Using Power Pivot to load web log data with Hive
3. Using Power View to analyze web log data

Big Data Modeling with Power Pivot

Benefits

Data models can surface big data in an intuitive way to promote rapid exploration, analysis and reporting

Big data can be easily integrated with other data sources

Self-service BI potential:

- Power Pivot can load big data by using the Table Import Wizard
  - ODBC direct to HDInsight
  - OLE DB with a SQL Server linked server to Azure HDInsight
- Power Pivot workbooks can become a data source for:
  - Local Excel reports (within the same workbook) with PivotTables, PivotCharts, CUBE functions and Power View
  - Other analytic and reporting tools (if published to SharePoint on-premises)

Big Data Modeling with Power Pivot

Considerations

Big data results may be too large for loading into in-memory storage

- Workarounds, by minimizing the amount of data to retrieve
  - Retrieve a smaller time period of data
  - Decrease the dimensionality, and/or
  - Increase the grain
  - Sample with a random distribution of data

Once the big data results are loaded (cached in memory), the data model can deliver high query performance

Summary

Big data refers to data sets so large and/or complex that they become awkward to work with in conventional ways

Hadoop can store petabytes of data reliably and execute huge distributed computations

- Big data query results often involve significant latency

Power BI includes authoring add-ins in Excel to query, analyze and visualize data sourced from Azure HDInsight

Resources

Microsoft Big Data web site

- http://www.microsoft.com/en-us/server-cloud/solutions/big-data.aspx

Azure HDInsight web site

- http://azure.microsoft.com/en-us/documentation/services/hdinsight/

Hortonworks tutorials

- http://hortonworks.com/tutorials
- Numerous tutorials are available to learn about big data by using the Hortonworks Sandbox

Thank You

Bitwise Solutions

people + technology + **training** = success