Big Data Analytics with Excel
Presenter Introduction

Peter Myers
Independent BI Expert – Bitwise Solutions
BBus, SQL Server MCSE, MCT, SQL Server MVP since 2007
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
LinkedIn: http://www.linkedin.com/in/peterjsmyers

Microsoft MVP (Most Valuable Professional)
Session Outline

Introducing:

- Big Data
- Hadoop
- Azure HDInsight
- Excel BI with “Power” Add-ins

Big Data Modeling with Power Pivot:

- Benefits
- Considerations

Resources
Introducing Big Data

- **Device Explosion**: > 5.5 billion (> 70% of global population)
- **Social Networks**: > 2 billion users
- **Ubiquitous Connection**: Web traffic
  - 2010: 130 exabyte (10^18)
  - 2015: 1.6 zettabyte (10^21)
- **Sensor Networks**: > 10 billion
- **Cheap Storage**: $100 gets you 3 million times more storage in 30 years
- **Inexpensive Computing**: 1980: 10 MIPS/$
  - 2005: 10M MIPS/$
“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)

Data Complexity: Variety and Velocity

- **Petabytes**
  - Click stream
  - Wikis/blogs
  - Advertising
  - Mobile
  - ERP/CRM
  - Payables
  - Payroll
  - Inventory

- **Terabytes**
  - Sensors
  - RFID
  - Social sentiment
  - Devices

- **Gigabytes**
  - Collaboration
  - eCommerce
  - Web Logs
  - Digital Marketing

- **Megabytes**
  - Web 2.0
  - Search Marketing
  - Recommendations
  - Weather
  - Text/image

- **Big Data**
  - Log files
  - Spatial and GPS coordinates
  - Data market feeds
  - eGov feeds
Introducing Big Data
Responding to New Questions

What’s the social sentiment of my product?

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

How do I better predict future outcomes?
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
Comparison to Traditional RDBMS

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

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)
- Develop in .NET and Java
- Can be automated with PowerShell and Command Line
- Can deliver insights through Excel with the “Power” add-ins
Introducing Azure HDInsight
How it Works: 1 – Data Storage
1. Storing Web Log Data in Azure Storage
Introducing Azure HDInsight
How it Works: 2 – Take the Processing to the Data
Introducing Azure HDInsight
The Big Data Ecosystem

Legend
Red = Core Hadoop
Blue = Data processing
Green = Packages
Purple = Microsoft integration points and value adds
Orange = Data Movement
Demonstrations

1. Word count with Pig ("Hello World" for Hadoop)
Introducing Azure HDInsight

Traditional E-Commerce Data Flow
Introducing Azure HDInsight
New E-Commerce Big Data Flow
Introducing Azure HDInsight
Hadoop Data Flow

Data  Hadoop  Analytics
Excel BI with “Power” Add-ins
Excel: Complete and Powerful Self-Service Tool

- Save to a Power BI site:
  - Browser access to workbooks, including:
    - Power Q&A
    - Forecasting
    - Data Catalog
    - Scheduled data refresh
    - Mobile application

- Power View
- Power Map

- Access
- Clean
- Mash-up

- Share

- Explore

- Power Query
- Power Pivot
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.

Excel BI includes “Power” add-ins to query, analyze and visualize Big Data sourced from Azure HDInsight.
Microsoft Big Data web site

Azure HDInsight web site

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