Big Data on the Microsoft Platform

With Hadoop, MS BI and the SQL Server stack

Andrew J. Brust, CEO, Blue Badge Insights
Meet Andrew

CEO and Founder, Blue Badge Insights

Big Data blogger for ZDNet
Microsoft Regional Director, MVP
Co-chair VSLive! and 17 years as a speaker
Founder, Microsoft BI User Group of NYC
  - http://www.msbigdatanyc.com
Co-moderator, NYC .NET Developers Group
  - http://www.nycdotnetdev.com
“Redmond Review” columnist for
Visual Studio Magazine
brustblog.com, Twitter: @andrewbrust
Read all about it!
My New Blog (bit.ly/bigondata)
Agenda

Big Data, Hadoop and HDInsight
MapReduce
Hive ODBC, BI Stack
Hekaton, NoSQL
SQL Server Parallel Data Warehouse, MPP, PolyBase
What is Big Data?

100s of TB into PB and higher
Involving data from: financial data, sensors, web logs, social media, etc.
Parallel processing often involved
  ▪ Hadoop is emblematic, but other technologies are Big Data too
Processing of data sets too large for transactional databases
  ▪ Analyzing *interactions*, rather than *transactions*
  ▪ The three V’s: Volume, Velocity, Variety
What is Hadoop?

Open source implementation of Google’s MapReduce and GFS (Google File System)

Allows for scale-out processing of petabyte scale data
- 1 PB = 1,024 TB

Also distributed storage

Commodity hardware

Can work against flat files, or certain database formats

Native processing involves imperative Java code

Other languages supported through “Streaming”
What is HDInsight?

Microsoft’s Hadoop distribution, on Windows
- Most other distros on Linux

Based on Hortonworks Data Platform (HDP)
Runs on Azure, eventually on Windows Server, and as sandbox on dev PC
For .NET devs: .NET SDK for Hadoop, LINQ provider
Demo
HDInsight
The Hadoop Stack

- Log file integration
- Machine Learning/Data Mining
- RDBMS Import/Export
- Query: HiveQL and Pig Latin
- Database
- MapReduce, HDFS
MapReduce, in a Diagram
A MapReduce Example

- Count by suite, on each floor
- Send per-suite, per platform totals to lobby
- Platform packets to 10th, 20th, 30th floor
- Tally up each floor
- Collect the tallies
- Merge tallies into one spreadsheet
MapReduce Options

- Java
- JavaScript (“Rhino”)
- Other languages, especially Python, via Streaming
- C# via Streaming
- C# via .NET SDK

Pig, Hive, Sqoop, Mahout also generate MapReduce code
Amenities for Visual Studio/.NET

- MR code in C#, HadoopJob, MapperBase, ReducerBase
- MRLib (NuGet Package)
- LINQ to Hive
- WebHDFS client
- OdbcClient + Hive ODBC Driver
- Deployment
- Hortonworks Data Platform for Windows
Hive

Began as Hadoop sub-project
  ▪ Now top-level Apache project

Provides a SQL-like ("HiveQL") abstraction over MapReduce

Has its own HDFS table file format (and it’s fully schema-bound)

Can also work over HBase

Acts as a bridge to many BI products which expect tabular data
Hive ODBC Consumers

- Excel 2010 or 2013 (including via add-in)
- PowerPivot
- SQL Server Analysis Services, Tabular Mode
- SQL Server Reporting Services
- ADO.NET OdbcClient provider
- LINQ provider
xVelocity Technologies

Formerly known as VertiPaq
PowerPivot, SSAS Tabular, SQL Server columnar indexes
Implements BI Semantic Model (BISM)
Uses column store technology
  - Compression
  - In-memory
  - Speed
Not a Big Data technology per se, but very useful for analysis of job output
Power View

Reports on BISM models (PowerPivot, SSAS Tabular)
Hosted in SharePoint 2010, 2013 Enterprise
Also Excel 2013 (but not on ARM/Windows RT)
Interactive data exploration
Demo
Hive ODBC + BI Stack
Project “Hekaton”

In-memory engine for SQL Server transactional workloads
Tables must be declared as in-memory explicitly
In-memory and standard tables can coexist in same db
Stored procs on in-mem tables are compiled to native code

Hekaton and xVelocity are separate
Hekaton ≠ PowerPivot/SSAS Tabular
Hekaton ≠ Columnstore indexes
Compare to SAP HANA
- In-memory, transactional, analytical, column store
NoSQL

NoSQL databases are non-relational and non- or loosely-schematized

HBase is a NoSQL database, of the wide column variety
- Hive implements a SQL layer over it
- HBase not yet in HDInsight

HBase table = HDFS file

Three other NoSQL categories
- Key-value store, document store, graph database
- Azure Table Storage is a key-value store NoSQL database

Some of them aren’t really Big Data tools, but market themselves that way anyway
SQL Parallel Data Warehouse (PDW)

SQL PDW is a Massively Parallel Processing (MPP) database
Teradata, IBM Netezza, HP Vertica also in this category
It’s an array/cluster of SQL servers made to look like one SQL Server
Available as appliance only
- Purchase from HP, Dell
- Server, storage and network all pre-built and configured

Many other MPP products based on PostgreSQL
PDW loosely based on acquired DATAAllegro product
- Implemented MPP with Ingres, written in Java, running on Linux
### MapReduce versus MPP

<table>
<thead>
<tr>
<th>MapReduce</th>
<th>MPP</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Splits preprocessing amongst mapper nodes and aggregation amongst reducers</td>
<td>- Splits query amongst nodes then unifies result sets</td>
</tr>
<tr>
<td>- Scales infinitely on commodity hardware</td>
<td>- Scales to high-end assets in the appliance cabinet</td>
</tr>
<tr>
<td>- Uses locally attached commodity disks on nodes</td>
<td>- Uses shared storage (can be more network traffic)</td>
</tr>
<tr>
<td>- Uses imperative code</td>
<td>- Uses SQL</td>
</tr>
<tr>
<td>- Processes flat files, wide column tables (HBase), relational tables (Hive)</td>
<td>- Works with relational tables only</td>
</tr>
<tr>
<td>- Divide-and-conquer approach, parallel, distributed</td>
<td>- Divide-and-conquer approach, parallel, distributed</td>
</tr>
</tbody>
</table>
PolyBase

To be included in next version of PDW
Mashup of SQL Server and Hadoop
Enables PDW to address Hadoop data nodes (HDFS) directly
Parallelism managed by PDW
Tables are imported into SQL Server db
- They are EXTERNAL tables
- They can participate in joins with standard tables
Resources

MS Big Data/HDInsight
- http://www.microsoft.com/bigdata

Apache Hadoop

Apache HBase
- http://hbase.apache.org/

SQL PDW

PolyBase
- http://gsl.azurewebsites.net/Projects/Polybase.aspx

xVelocity

Column store

Power View

Hekaton
Questions?
Thank You for Attending