AZURE HDINSIGHT

Azure Machine Learning Track
Marek Chmel
SESSION AGENDA

- Understanding different scenarios of Hadoop
- Building an end to end pipeline using HDInsight
- Using in-memory techniques to analyze data interactively
BIG DATA VS. TRADITIONAL DW
TWO APPROACHES TO INFORMATION MANAGEMENT FOR ANALYTICS: TOP-DOWN + BOTTOM-UP

Top-Down (Deductive)

- Theory
- Hypothesis
- Observation
- Confirmation

Bottom-Up (Inductive)

- Theory
- Hypothesis
- Pattern
- Observation

DIFFICULTY

INFORMATION

VALUE

OPTIMIZATION

What happened?
Descriptive Analytics

Why did it happen?
Diagnostic Analytics

What will happen?
Predictive Analytics

How can we make it happen?
Prescriptive Analytics
DATA WAREHOUSING USES A TOP-DOWN APPROACH
THE “DATA LAKE” USES A BOTTOM-UP APPROACH

Ingest all data regardless of requirements

Store all data in native format without schema definition

Do analysis Using analytic engines like Hadoop

- Batch queries
- Interactive queries
- Real-time analytics
- Machine Learning
- Data warehouse
What happened?
What is happening?
Why did it happen?
What are key relationships?

What will happen?
What if?
How risky is it?
What should happen?
What is the best option?
How can I optimize?
WHAT IS HDINSIGHT
MICROSOFT HADOOP STACK

Analytics

- Azure HDInsight
  - Hadoop
  - Interactive
  - Hive
  - HBase
  - NoSql
  - Storm
  - Real Time
  - Spark
  - Streaming
  - Interactive
  - Batch
  - ML
  - R Server
  - Machine Learning

Storage

- Local (HDFS) or Cloud (Azure Blob/Azure Data Lake Store)

Hadoop Distributions running in Azure VMs

- Hortonworks
- Cloudera
- MapR
Hadoop clusters have grown by 60% in the last 2 years

89% of enterprise users consider Hadoop as opportunity for innovation

Forrester report predicts that Hadoop will grow by 33% annually in next five years

Hadoop is shifting from a buzzword to a real production service

Ownership is shifting from department teams to Central IT.
AZURE HDINSIGHT

- Fully-managed Hadoop and Spark for the cloud
- 100% Open Source Hortonworks data platform
- Clusters up and running in minutes
- Supported by Microsoft with industry’s best SLA
- Familiar BI tools for analysis
- Open source notebooks for interactive data science
- 63% lower TCO than deploying Hadoop on-premise*

Hadoop and Spark as a Service on Azure
HDINSIGHT WORKLOADS

- Hadoop
  - Batch: Hive and MapReduce
  - Interactive Hive using LLAP (New – just launched)

- HBase (NoSQL)

- Storm (Streaming)

- Spark (Interactive)
INTRO TO HIVE

Platform
- ETL
- Reporting
- Data Mining
- Deep Analytics

Core SQL Engine
- Sub-Second SQL
- Ad-Hoc
- Drill-Down
- BI Tools: Tableau, Excel

Connectivity
- Continuous ingestion from operational DB
- Slowly changing dimensions

Legend
- Existing
- Development
- Emerging

High Perf Batch SQL
- Interactive SQL
- Reporting
  - BI Tools: Tableau, Excel etc.

Multidimensional Analytics
- MDX Tools
  - Excel

AcID/Merge
- OLAP/Cube
  - MDX Tools
  - Excel

Compute
- SQL 2011 Compiler

Storage
- Cost based optimizer

Platform
- Tez Execution Engine

Core Hive
- Security
CREATING HDINSIGHT CLUSTER
CREATING HDINSIGHT CLUSTER
CLUSTER DASHBOARD

- Ambari dashboard which you can use with Hortonworks HDP
TYPICAL HADOOP SCENARIOS

- **ETL**
  - Data ingested from various sources
  - Transformed and cooked to structured data
  - Then loaded into a DB for querying
  - Typically batch scenario

- **BI Scenarios**
  - Used by business analyst for ad-hoc querying
  - Requires interactive response
BUILDING AN ENTERPRISE DW USING HADOOP

- Planning
  - Cluster planning
  - Cluster Deployment model

- Development
  - Author and Debug Queries
  - Optimize queries

- Deployment
  - Use ADF/Oozie to schedule and productionalize your jobs
  - Monitor and manager cluster using Ambari

- Connecting with BI tools
  - Create tables on ORC data from shared storage account
  - Have BI tools connect to cluster using ODBC driver
CLUSTER PLANNING

• Understand requirements
  • What is scenario?
  • What is SLA?
  • What is budget?
  • How often?
  • Who is the customer?

• Type of cluster
  • Production, Dev or Test?
  • On-demand vs. persistent?
  • Custom vs. default metastore?
  • Security model?

• Trade-offs
  • Single or multi tenant?
  • CPU or Memory bound?
## CLOUD DEPLOYMENT MODELS

<table>
<thead>
<tr>
<th></th>
<th>Always on cluster (Persistent)</th>
<th>Cluster as a service (On demand)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Storage choice</td>
<td>Local HDFS, Azure Blob, Azure Data Lake Store</td>
<td>Azure Blob, Azure Data Lake Store</td>
</tr>
<tr>
<td>Job Scheduling</td>
<td>Oozie</td>
<td>Azure Data Factory</td>
</tr>
<tr>
<td>Data persistence after cluster deletion</td>
<td>N/A</td>
<td>Azure Blob, Azure Data Lake Store</td>
</tr>
<tr>
<td>Metadata persistence after cluster deletion</td>
<td>N/A</td>
<td>Azure SQL</td>
</tr>
<tr>
<td>Billing</td>
<td>Billing for entire time cluster is up</td>
<td>Billing per job</td>
</tr>
</tbody>
</table>

**Why use Cluster as a Service?**
- Pay only for time the cluster was actually used
- Since data & metadata is persisted, experience is as if the cluster was never deleted
Ambari Views
- Provides graphical UX for authoring and debugging Hive queries
- Pros: One of the few tools that can be used to debug Tez queries

Visual Studio
- Enables writing Hive queries using Visual Studio
- Pros: Offers choice between Templeton and HiveServer2

Command-Line
- Provides SSH and Windows CLI access
- Pros (and also cons): Very powerful

Beeline
- Command line shell that works with HiveServer2.
- Pros: Very thin JDBC client
INTRODUCING HIVE LLAP: MAKING HIVE INTERACTIVE
HIVE LLAP

- Interactive Querying through in-memory compute
- 10x-25x faster than using Hive
- Allows multiple users to run queries simultaneously
- Provides enterprise class security
- Separate capacity for ETL and EDW scenarios
- Integration with world class BI tools
SCALING FOR BIG DATA WORKLOADS

- Challenges
  - Improving High Availability
  - Elastic scaling
  - Ability to scale to multiple users

- How HDInsight helps with scaling
  - Platform Availability improvements
  - Ability to scale during and after cluster creation
  - Ability to create Edge Nodes
HDINSIGHT SECURITY – RINGS OF DEFENSE

Perimeter Level Security
- Virtual Network
- Network Security (i.e. Firewalls)
- Gateway

Authentication
- Kerberos
- Active Directory

Authorization
- Hive policies
- HBase policies
- File and Folder level ACLS

Data Security
- Encryption @ Rest
INTEGRATION WITH AZURE ACTIVE DIRECTORY

Authentication
Kerberos
Active Directory
APPLICATION AND DATA-LEVEL AUTHORIZATION

Authorization
Hive policies
HBase policies
File and Folder level ACLS
SECURE ENDPOINTS IN HDINSIGHT CLUSTER

Access to all users
- HiveServer2
- Ambari & Views
- Ranger

Access to only Cluster Admin
- SSH
- WebHCat
- Oozie
Azure Data Lake Storage
- Public Preview
- ALWAYS ON transparent encryption
- All reads/writes are encrypted/decrypted
- Service managed keys as well as Customer managed keys

Windows Azure Storage Blob
- General Availability
- ALWAYS ON transparent encryption
- All reads/writes are encrypted/decrypted
- Service managed keys