Data Analytics using Microsoft's Azure Cloud

Azure Batch, Azure Data Lake, Azure HDInsight, ML, Power BI

Power BI Meetup
April 13, 2017

Roy Kim
@RoyKimYYZ
roykimtoronto@gmail.com
Agenda

- Overview of Big Data + Azure + Data Insights
- Job Postings demo solution architecture & implementation
- Mobile Demo with Power BI
- Q&A
Roy Kim

14+ Years of Microsoft Technology Solutions

.NET, SharePoint, BI, Office 365, Azure Solutions

Currently an IT Consultant

University of Toronto – Computer Science Degree
Data to Insight

Big Data  Data Platform Technologies  Solution  Data Insights

By: Roy Kim
Job Postings Demo Solution

Job Postings
Azure Data Platform
Data Lake, HDInsight, SQL, Power BI
Job trends, analysis

By: Roy Kim
Figure 1. SCM Data Volume and Velocity vs. Variety

References:
https://softwarestrategiesblog.com/2015/09/05/10-ways-big-data-is-revolutionizing-supply-chain-management
Many services growing and maturing
Azure Data Platform

Two Illustrations:

Data models and analytics
- SQL Server Analysis Services
- Multidimensional and Tabular data models
- Data mining

Data stores
- SQL Server
- Azure storage
- Hadoop platform
- HBase
- Azure SQL Database
- SQL Server Parallel Data Warehouse

Integration
- SQL Server Integration Services
- Azure Intelligent Systems Service
- StreamInsight
- Data Quality Services

Data sources
- Business applications
- Master Data Services
- Device sensors & streaming data feeds
- Clickstream data & web logs

Visualization and reporting
- Power BI for Office 365
- SharePoint Server

Users
- Web (HTML)
- Mobile phones
- Tablets

Collaboration
- Office 365
- Ad-hoc
- Power BI Reporting Services IaaS

Visualizations
- Power BI
- Power BI

Real-time Analytics
- Azure Machine Learning
- Analysis Services IaaS

Predictive Analytics
- HortonWorks Data Platform (HDP) on Windows Server
- HDInsight on Microsoft Azure

Historical Analytics
- Analytics Platform System (APS) on Parallel Data Warehouse (PDW)

By: Roy Kim

https://blogs.technet.microsoft.com/cansql/2015/06/03/microsoft-data-platform-overview/
Figure 1. Magic Quadrant for Business Intelligence and Analytics Platforms

Source: Gartner (February 2017)
The Analytics Continuum

**Data**

- **Analytics**
  - *Descriptive*
    - What happened?
  - *Diagnostic*
    - Why did it happen?
  - *Predictive*
    - What will happen?
  - *Prescriptive*
    - What should I do?

**Human Input**

- *Decision*
- *Action*

- *Decision Support*
- *Decision Automation*

Feedback
Job Postings Data Set

**Volume**
- Many national job sites
- New job postings daily
- Metadata and full text.

**Velocity**
- New job postings created every minute

**Variety**
- Semi-structured
  - Job Title
  - Location
  - Company
- Unstructured
  - Job Description

**Veracity**
- Incomplete/Imprecise
- Salary, Per hour
- FT, PT, Temp, Contract, Seasonal
- Main profession

By: Roy Kim
Power BI – Job Postings Demo Reports

By: Roy Kim
Job Postings Big Data Solution Architecture

Presentation Tier

Visualization/Reporting Tools

Services Tier

Storage Tier

By: Roy Kim
Job Postings from Internet Job Boards

- Web sites that offer APIs such as indeed.com, dice.com, etc.
- Use any server-side programming language to retrieve data such as NET, Java, Node.js, etc.
- If no APIs, consider HTML web page scraping

REST API
- http end points typically return JSON or XML data formats

Html Web Page Scraping

- HTML Agility Pack to assist in parsing the Document Object Model for data points.
  - [https://www.nuget.org/packages/HtmlAgilityPack](https://www.nuget.org/packages/HtmlAgilityPack)
  - HTML parsing supporting XPath to traverse the Document Object Model (DOM)
  - E.g. doc.DocumentElement.SelectSingleNode("//div[@id='Total Sales']")

By: Roy Kim
Azure Application Insights

Application Insights Core API. This package provides core functionality for transmission of all Application Insights Telemetry Types and is a dependent package for all other Application Insights packages.
A managed Azure service executing command line applications.

For batch processing or batch computing--running a large volume of similar tasks to get some desired result.

Commonly used by organizations that regularly process, transform, and analyze large volumes of data.

Simply, a set of Azure Virtual Machines running a console application to process data that can be on a recurring schedule and in parallel.

References:
https://github.com/Microsoft/azure-docs/blob/master/articles/batch/batch-technical-overview.md

Author: Roy Kim
Azure Data Lake

- Intended for data storage in its raw format for future analysis, processing or data modelling.
- For developers, data scientists, and analysts to store data of any size, shape, and speed.
- To do all types of processing and analytics across different platforms and languages.
- Extract and load, minimal transformations
- To manage data in characteristic of variety, velocity and volume

Two Components
  1. Azure Data Lake Store
  2. Azure Data Lake Analytics

Azure Data Lake Store

- Azure Data Lake Store is a hyper-scale repository for big data analytic workloads. Azure Data Lake enables you to capture data of any size, type, and ingestion speed in one single place for operational and exploratory analytics.
- The Azure Data Lake store is an Apache Hadoop file system compatible with Hadoop Distributed File System (HDFS)
- Can be accessed from Hadoop (available with HDInsight cluster) using the WebHDFS-compatible REST APIs

References:
Use Cases

- Store social media posts, log files, sensor data
- Store corporate data such as relational databases (as flat files)

References:

By: Roy Kim
Azure Data Lake Analytics

- Azure Data Lake Analytics is built to make big data analytics easy.
- Focus on writing, running, and managing jobs, rather than operating distributed infrastructure. Instead of deploying, configuring, and tuning hardware.
- Write queries to transform your data and extract valuable insights. The analytics service can handle jobs of any scale instantly by setting the dial for how much power you need.
  - U-SQL – a Big Data query language. Likeness of SQL + C#
  - ”schema on reads”
- Pay for your job when it is running; making it cost-effective.

- Data Collector app stores .json files in respective folders
- USQL scripts logic:
  - reads 1000s of JSON files in a given folder
  - Outputs to one TSV (tab delimited) file
  - Create a Tables to schematize the TSV files
  - Query against tables to analyze or transform to a new output file.

References:

By: Roy Kim
Azure Data Lake Analytics – Demo Implementation

- USQL script: process json files into a tab delimited file

```usql
REFERENCE ASSEMBLY [Newtonsoft.Json];
REFERENCE ASSEMBLY [Microsoft.Analytics.Samples.Formats];

DECLARE @inputfile string="/jobpostings/-{}.json";

@jobPostingsSchema =
EXTRACT jobtitle string,
   company string,
   city string,
   state string,
   country string,
   formattedLocation string,
   date string,
   snippet string,
   url string,
   latitude float, longitude float, jobkey string, sponsored string, expired string,
   formattedLocationFull string,
   stations string,
   jobDescription string,
   salaryRate string,
   salaryType string,
   jobType string
FROM @inputfile

OUTPUT @jobPostingsSchema
TO "/jobpostings/outputtsv/v3/JobsOutput.tsv"
USING Outputters.Tsv();
```
Hadoop refers to an ecosystem of open-source software that is a framework for distributed processing, storing, and analysis of big data sets on clusters of commodity computer hardware.

Azure HDInsight makes the Hadoop components from the Hortonworks Data Platform (HDP) distribution available in Azure, deploys managed clusters with high reliability and availability, and provides enterprise-grade security and governance with Active Directory.

HDInsight offers the cluster types - Hadoop, HBase, Spark, Kafka, Interactive Hive, Storm, customized, etc.

Supports integration with BI tools such as Power BI, Excel, SQL Server Analysis Services, and SQL Server Reporting Services.

By: Roy Kim
Considerations

- To manage the compute costs, script the provisioning and de-provisioning of the cluster.
- While a cluster is running, execute scripts and query the data into self service BI tools and into other data warehouses.
- In comparison to Azure Data Lake, ADL Analytics may be more cost effective since it is pay per use at a more granular level - # of nodes and execution time. E.g. Running against 100 nodes may cost a few dollars per minute in ADL Analytics; whereas, in HDInsight, 13 nodes for small VM size may cost a few dollars an hour.
A relational database-as-a-service in the cloud built on the Microsoft SQL Server engine
- No need to manage the infrastructure.
- Scale up or down based on Database Transaction Units (DTUs).
- 1TB storage maximum
- Can be used as a simpler data warehouse.
- Developed a simple data warehouse modelling
- Job Postings data loaded from ADLS
- Star schema
- Added a date dimension table
- Table of # of jobs for each province by a date hierarchy

### Table

<table>
<thead>
<tr>
<th>Year</th>
<th>MonthName</th>
<th>AB</th>
<th>BC</th>
<th>NS</th>
<th>ON</th>
<th>QC</th>
<th>SK</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>2016</td>
<td>December</td>
<td>67</td>
<td>15</td>
<td>42</td>
<td>62</td>
<td>76</td>
<td>15</td>
<td>277</td>
</tr>
<tr>
<td></td>
<td></td>
<td>21</td>
<td>6</td>
<td>6</td>
<td>25</td>
<td>27</td>
<td>5</td>
<td>90</td>
</tr>
<tr>
<td></td>
<td></td>
<td>87</td>
<td>21</td>
<td>41</td>
<td>72</td>
<td>52</td>
<td>16</td>
<td>289</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>175</td>
<td>42</td>
<td>89</td>
<td>159</td>
<td>155</td>
<td>36</td>
<td>656</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Year</th>
<th>MonthName</th>
<th>AB</th>
<th>BC</th>
<th>NS</th>
<th>ON</th>
<th>QC</th>
<th>SK</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>43</td>
<td>23</td>
<td>21</td>
<td>18</td>
<td>80</td>
<td>46</td>
<td>231</td>
</tr>
<tr>
<td></td>
<td></td>
<td>58</td>
<td>12</td>
<td>24</td>
<td>39</td>
<td>74</td>
<td>19</td>
<td>226</td>
</tr>
<tr>
<td></td>
<td></td>
<td>42</td>
<td>10</td>
<td>7</td>
<td>11</td>
<td>34</td>
<td>17</td>
<td>121</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7</td>
<td>8</td>
<td>5</td>
<td>6</td>
<td>43</td>
<td>2</td>
<td>71</td>
</tr>
<tr>
<td></td>
<td></td>
<td>95</td>
<td>46</td>
<td>31</td>
<td>35</td>
<td>92</td>
<td>8</td>
<td>307</td>
</tr>
<tr>
<td></td>
<td></td>
<td>83</td>
<td>361</td>
<td>21</td>
<td>100</td>
<td>79</td>
<td>14</td>
<td>658</td>
</tr>
<tr>
<td></td>
<td></td>
<td>153</td>
<td>27</td>
<td>44</td>
<td>117</td>
<td>67</td>
<td>2</td>
<td>410</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>481</td>
<td>487</td>
<td>153</td>
<td>326</td>
<td>469</td>
<td>108</td>
<td>2024</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Year</th>
<th>MonthName</th>
<th>AB</th>
<th>BC</th>
<th>NS</th>
<th>ON</th>
<th>QC</th>
<th>SK</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>87</td>
<td>44</td>
<td>31</td>
<td>35</td>
<td>118</td>
<td>45</td>
<td>360</td>
</tr>
<tr>
<td></td>
<td></td>
<td>127</td>
<td>6</td>
<td>79</td>
<td>283</td>
<td>97</td>
<td>31</td>
<td>623</td>
</tr>
<tr>
<td></td>
<td></td>
<td>80</td>
<td>24</td>
<td>17</td>
<td>14</td>
<td>48</td>
<td>8</td>
<td>191</td>
</tr>
<tr>
<td></td>
<td></td>
<td>9</td>
<td>6</td>
<td>3</td>
<td>8</td>
<td>27</td>
<td>7</td>
<td>60</td>
</tr>
<tr>
<td></td>
<td></td>
<td>86</td>
<td>29</td>
<td>47</td>
<td>26</td>
<td>130</td>
<td>25</td>
<td>345</td>
</tr>
<tr>
<td></td>
<td></td>
<td>81</td>
<td>48</td>
<td>36</td>
<td>31</td>
<td>112</td>
<td>33</td>
<td>341</td>
</tr>
<tr>
<td></td>
<td></td>
<td>27</td>
<td>53</td>
<td>33</td>
<td>27</td>
<td>113</td>
<td>29</td>
<td>282</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>497</td>
<td>210</td>
<td>246</td>
<td>426</td>
<td>645</td>
<td>178</td>
<td>2202</td>
</tr>
</tbody>
</table>
Azure Data Factory

- Cloud-based data integration service that orchestrates and automates the movement and transformation of data.
- Create data pipelines that move and transform data, and then run the pipelines on a specified schedule (hourly, daily, weekly, etc.)
# Azure Data Factory

<table>
<thead>
<tr>
<th>Category</th>
<th>Data store</th>
<th>Supported as a source</th>
<th>Supported as a sink</th>
</tr>
</thead>
<tbody>
<tr>
<td>Azure</td>
<td>Azure Blob storage</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Azure Data Lake Store</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Azure SQL Database</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Azure SQL Data Warehouse</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Azure Table storage</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Azure DocumentDB</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Azure Search Index</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Databases</td>
<td>SQL Server*</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Oracle*</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>MySQL*</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>DB2*</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Teradata*</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>PostgreSQL*</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Sybase*</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Cassandra*</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>MongoDB*</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Amazon Redshift</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>File</td>
<td>File System*</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>HDFS*</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Amazon S3</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>FTP</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Others</td>
<td>Salesforce</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Generic ODBC*</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Generic OData</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Web Table (table from HTML)</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>GE Historian*</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

By: Roy Kim
Predicting Salary for a given set of parameters such as job title and location
Azure Machine Learning – Demo Implementation

JobPostings Salary Experiment → Score Model → Scored dataset

<table>
<thead>
<tr>
<th>Jobtitle</th>
<th>State</th>
<th>Salary</th>
<th>KeyPhrase</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dynamic Sales Administrator</td>
<td>BC</td>
<td>32500</td>
<td>environment of extreme customer service, advanced computer skills, service standards, retail environment, team members, proven background of success, highly skilled team of high achievers, advanced levels of productivity, advanced levels of creativity, world challenging fast-paced environment of uncompromising...</td>
</tr>
<tr>
<td>Administrators, Department of Psychology and Neuroscience, ADM07</td>
<td>NS</td>
<td>65463</td>
<td>Dalhousie University, management, human resources, management, university, employee, university community, Neuroscience Department, international university...</td>
</tr>
</tbody>
</table>

Statistics
- Mean: 54916.7451
- Median: 45000
- Min: 15000
- Max: 350000
- Standard Deviation: 39275.3156
- Unique Values: 248
- Missing Values: 0
- Feature Type: Numeric Label

Visualizations
- ScatterPlot
  - compare to: Scored Labels
The main features of your Power BI service UI:
1. navigation bar
2. dashboard with tiles
3. Q&A question box
4. help and feedback buttons
5. dashboard title
6. Office 365 app launcher
7. Power BI home buttons
8. Additional dashboard actions
Power BI Mobile

By: Roy Kim
Key Mobile Scenarios

- Frequently updated and accessed reports
  - Minutes, hours, daily, weekly
- Fast and easy access of reports and dashboards
- IoT and sensor data
- Retail and customer analytics
- Team and organizational performance and productivity e.g. ticket management
- Collaborative analysis and decision making
- Not always in front of a large screen device

By: Roy Kim
Mobile App IOS Key Features & Demo

- Navigation
- Dashboards and Reports
- Responsive design
- Visualization interaction
- Sharing
- Annotations
- Q&A
- Alerts
- Favourites
Cloud services such as Azure Data Platform provide new capabilities in Data Analytics. That is in terms of scale, cost and agility.

Azure Data Lake is a productive option for organizations new to Hadoop. Yet continue to plan for other Hadoop offerings best fit for other scenarios.

Many azure services fit together to make the appropriate solution. That is SaaS, PaaS, IaaS, Data, App, Operational, etc.

As part of planning and design, be aware of MS roadmap and industry trends.
Appendix - Architecture

- Security
- Capacity
- Operations
- Application Design
- Business Processes
- Performance
- Data

Governance