

# 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

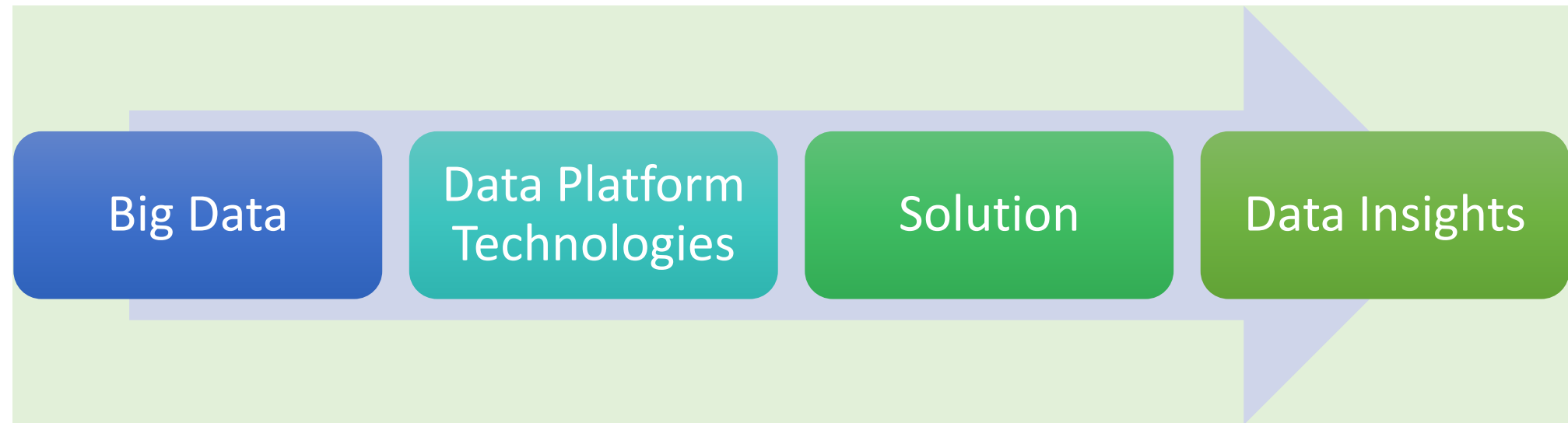
# Agenda

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

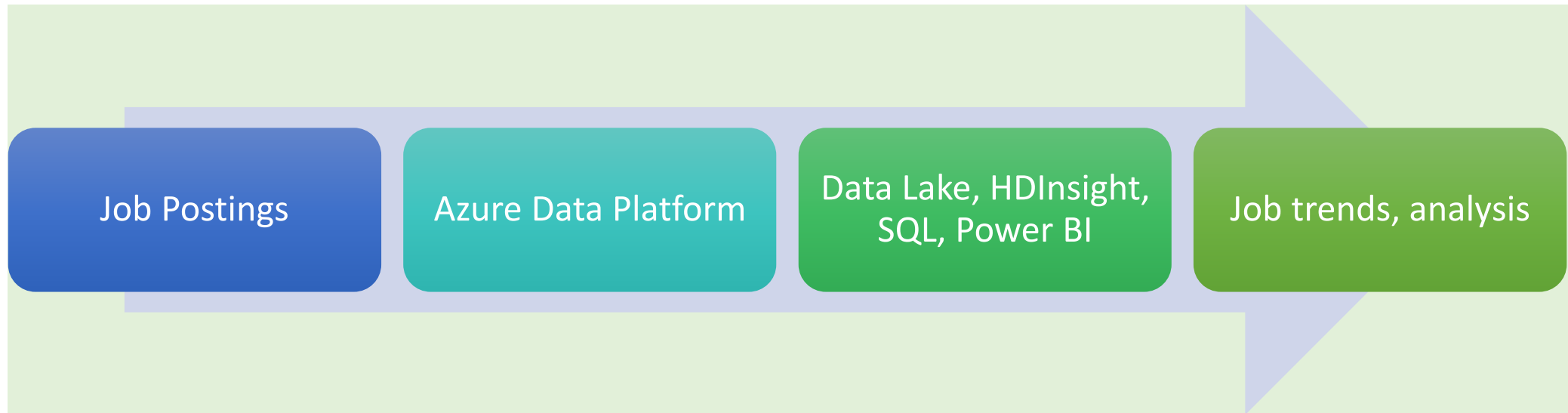
# Bio

- 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

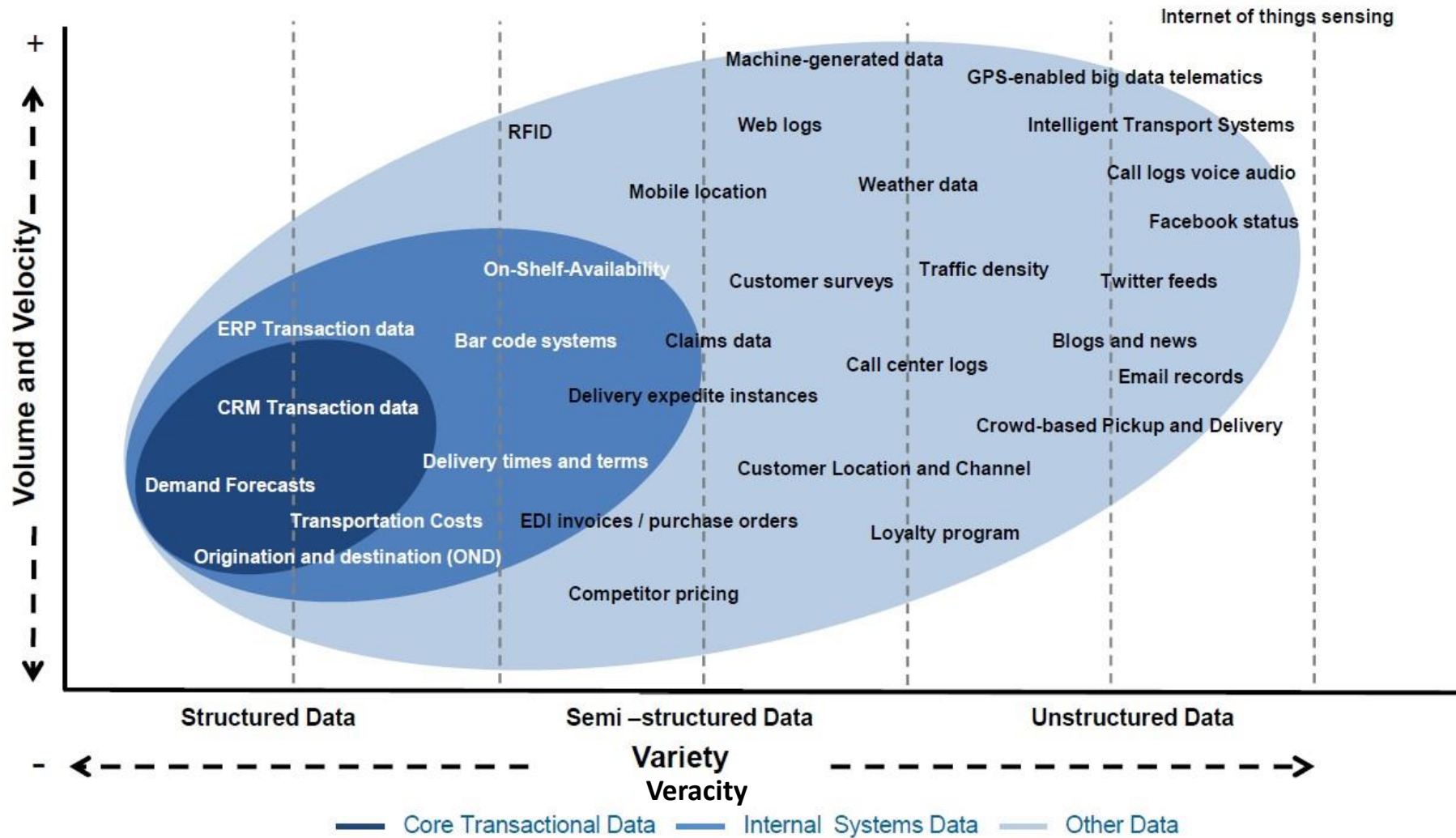
# Data to Insight



# Job Postings Demo Solution



# Big Data Spectrum



**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>

# Azure Cloud Platform

Many services growing and maturing

## What is Microsoft Azure?

### Virtual Machines

When you build cloud building blocks, you get control over a virtual machine or self-managed platform. Virtual machines are used to run applications and services. They are managed by the operating system, and you can manage them using the Azure portal or the Azure CLI.

### Cloud Services

Cloud services are managed by the Azure platform. They are managed by the Azure platform, and you can manage them using the Azure portal or the Azure CLI.

### App Service

App Service is a high-productivity solution for developing and running web applications. It is managed by the Azure platform, and you can manage it using the Azure portal or the Azure CLI.

## Microsoft Azure

Microsoft Azure is a flexible, open, and secure public cloud built for business. Access a broad collection of integrated services that accommodate many languages and operating systems. Use world-class tools to accelerate a wide variety of app development and delivery capabilities.

**Free trial!**

Use \$200 credit to try any combination of Azure resources.

[aka.ms/TryAzure](https://aka.ms/TryAzure)

Search Azure, Microsoft.com, MSDN, or TechNet for keywords used in this guide.

## Catalog of Services

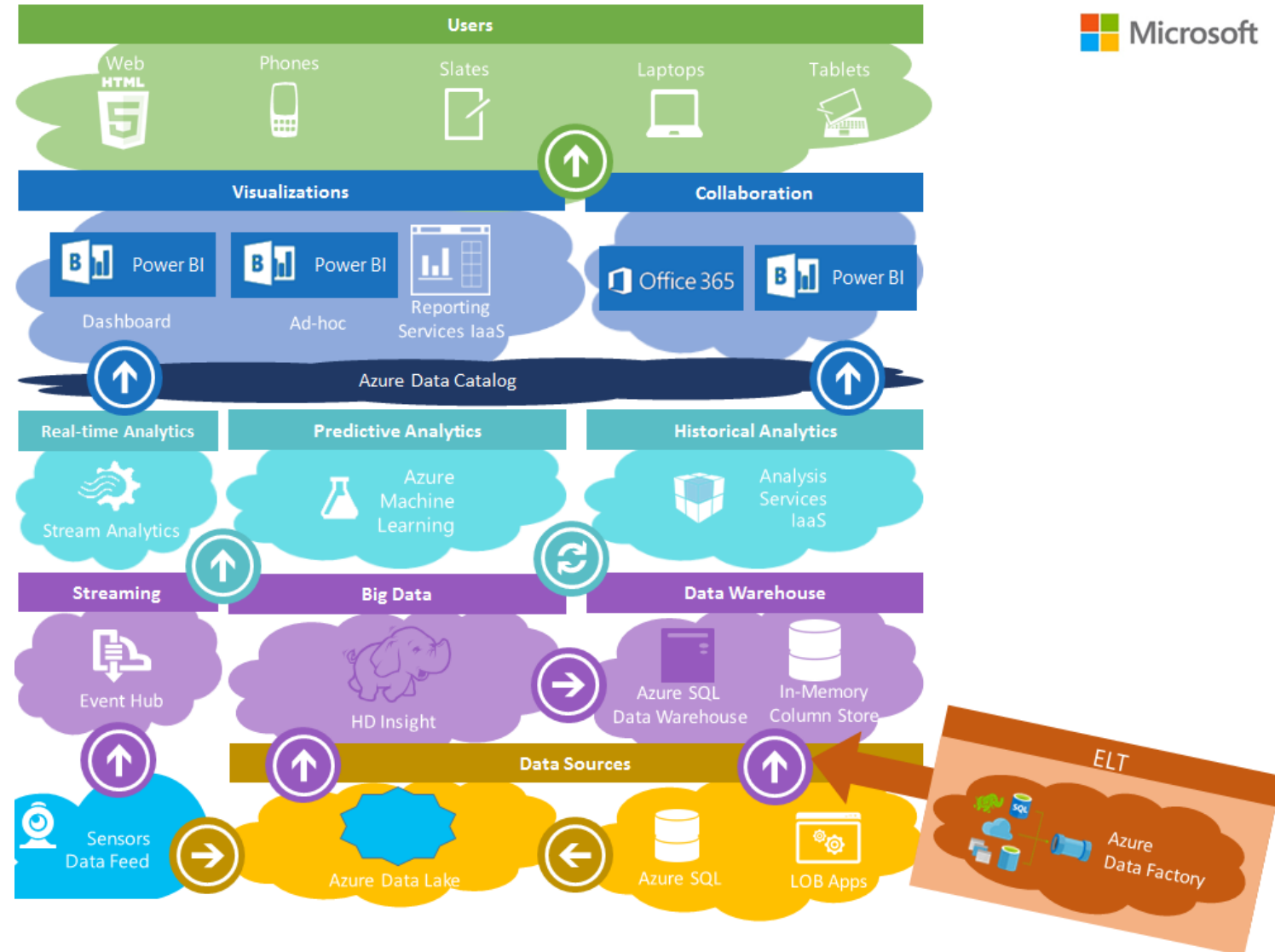
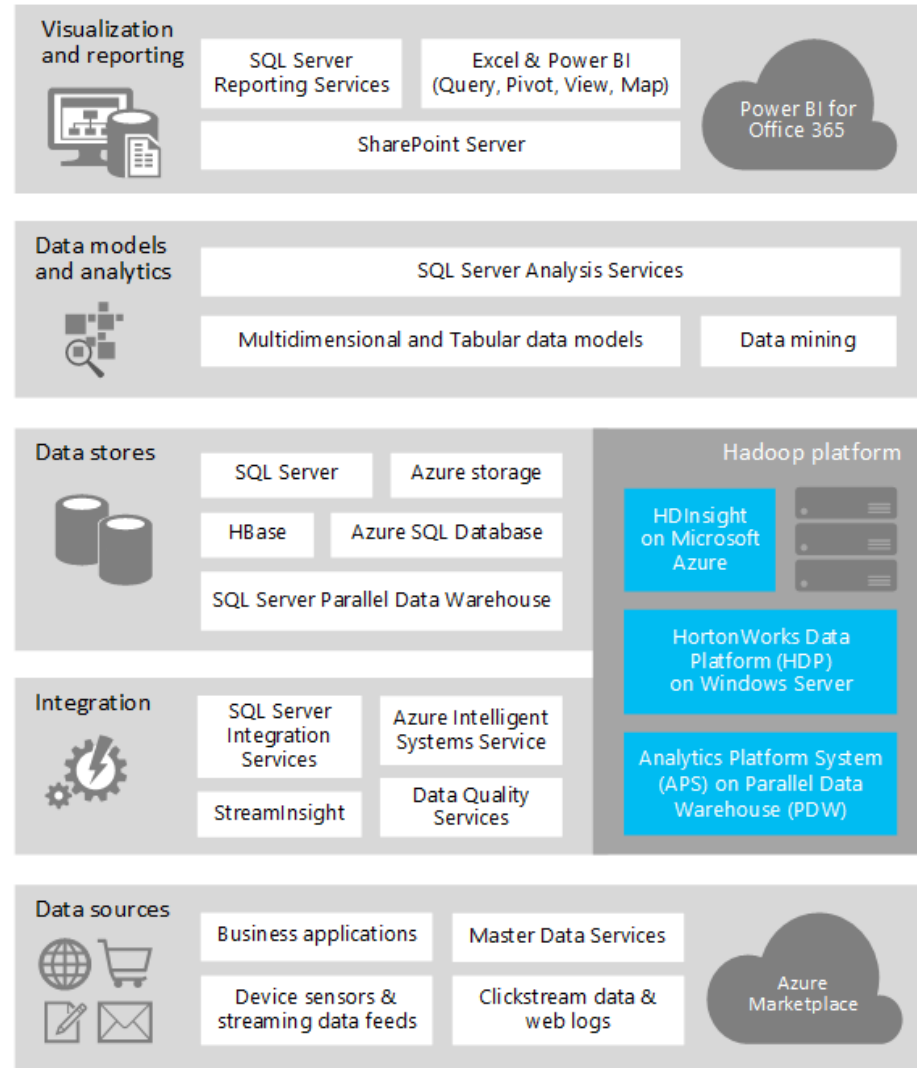
COMPUTE	WEB & MOBILE	STORAGE & BACKUP	HYBRID INTEGRATION	NETWORKING	ANALYTICS	IDENTITY & ACCESS	DEVELOPER SERVICES	MEDIA & CDN
<b>Virtual Machines</b> Run any workload on a server in the cloud.	<b>Web Apps</b> Manage web applications in the cloud.	<b>Storage Blobs &amp; Files</b> Store any type of data in the cloud.	<b>Storage Queues</b> Send messages between applications.	<b>Virtual Network</b> Manage your network in the cloud.	<b>HDInsight</b> Run Hadoop and other analytics in the cloud.	<b>Active Directory</b> Manage your identity in the cloud.	<b>Visual Studio Online</b> Develop and test your applications in the cloud.	<b>Media Services</b> Store and manage your media in the cloud.
<b>Cloud Services</b> Managed cloud services that you can use to build your applications.	<b>Mobile Apps</b> Build and manage mobile applications in the cloud.	<b>Backup</b> Backup your data in the cloud.	<b>BizTalk Services</b> Integrate your on-premises applications with the cloud.	<b>Express Route</b> Connect your on-premises network to the cloud.	<b>Machine Learning</b> Build and train machine learning models in the cloud.	<b>Multi-Factor Authentication</b> Add an extra layer of security to your applications.	<b>App Service</b> Build and run web applications in the cloud.	<b>CDN</b> Deliver content to your users in the cloud.
<b>Service Fabric</b> Build and manage distributed applications in the cloud.	<b>API Apps</b> Build and manage API applications in the cloud.	<b>Import / Export</b> Import and export data to and from the cloud.	<b>Hybrid Connections</b> Connect your on-premises applications to the cloud.	<b>Traffic Manager</b> Route traffic to the cloud.	<b>Stream Analytics</b> Process and analyze streaming data in the cloud.	<b>Event Hubs</b> Collect and analyze events in the cloud.	<b>App Service</b> Build and run web applications in the cloud.	<b>CDN</b> Deliver content to your users in the cloud.
<b>Batch</b> Run batch workloads in the cloud.	<b>Logic Apps</b> Build and manage logic applications in the cloud.	<b>Site Recovery</b> Recover your data in the cloud.	<b>Service Bus</b> Send messages between applications in the cloud.	<b>Remote App</b> Run your applications in the cloud.	<b>DocumentDB</b> Store and query documents in the cloud.	<b>Search</b> Search and index data in the cloud.	<b>App Service</b> Build and run web applications in the cloud.	<b>CDN</b> Deliver content to your users in the cloud.
<b>Scheduler</b> Run scheduled tasks in the cloud.	<b>API Management</b> Manage your API applications in the cloud.	<b>StorSimple</b> Store and manage your data in the cloud.	<b>Automation</b> Automate your tasks in the cloud.	<b>Virtual Network</b> Manage your network in the cloud.	<b>SQL Database</b> Store and query data in the cloud.	<b>Tables</b> Store and query data in the cloud.	<b>App Service</b> Build and run web applications in the cloud.	<b>CDN</b> Deliver content to your users in the cloud.
<b>Remote App</b> Run your applications in the cloud.	<b>Notification Hubs</b> Send notifications to your users in the cloud.	<b>SQL Database</b> Store and query data in the cloud.	<b>Portal</b> Manage your resources in the cloud.	<b>Express Route</b> Connect your on-premises network to the cloud.	<b>DocumentDB</b> Store and query documents in the cloud.	<b>Search</b> Search and index data in the cloud.	<b>App Service</b> Build and run web applications in the cloud.	<b>CDN</b> Deliver content to your users in the cloud.
<b>Batch</b> Run batch workloads in the cloud.	<b>API Management</b> Manage your API applications in the cloud.	<b>StorSimple</b> Store and manage your data in the cloud.	<b>Automation</b> Automate your tasks in the cloud.	<b>Traffic Manager</b> Route traffic to the cloud.	<b>SQL Database</b> Store and query data in the cloud.	<b>Tables</b> Store and query data in the cloud.	<b>App Service</b> Build and run web applications in the cloud.	<b>CDN</b> Deliver content to your users in the cloud.
<b>Scheduler</b> Run scheduled tasks in the cloud.	<b>Notification Hubs</b> Send notifications to your users in the cloud.	<b>SQL Database</b> Store and query data in the cloud.	<b>Automation</b> Automate your tasks in the cloud.	<b>Express Route</b> Connect your on-premises network to the cloud.	<b>DocumentDB</b> Store and query documents in the cloud.	<b>Search</b> Search and index data in the cloud.	<b>App Service</b> Build and run web applications in the cloud.	<b>CDN</b> Deliver content to your users in the cloud.
<b>Remote App</b> Run your applications in the cloud.	<b>Notification Hubs</b> Send notifications to your users in the cloud.	<b>SQL Database</b> Store and query data in the cloud.	<b>Automation</b> Automate your tasks in the cloud.	<b>Traffic Manager</b> Route traffic to the cloud.	<b>SQL Database</b> Store and query data in the cloud.	<b>Tables</b> Store and query data in the cloud.	<b>App Service</b> Build and run web applications in the cloud.	<b>CDN</b> Deliver content to your users in the cloud.

References:

<https://blogs.technet.microsoft.com/cansql/2015/06/03/microsoft-data-platform-overview/>

# Azure Data Platform

## Two Illustrations:





# Analytics Platform Gartner Magic Quadrant

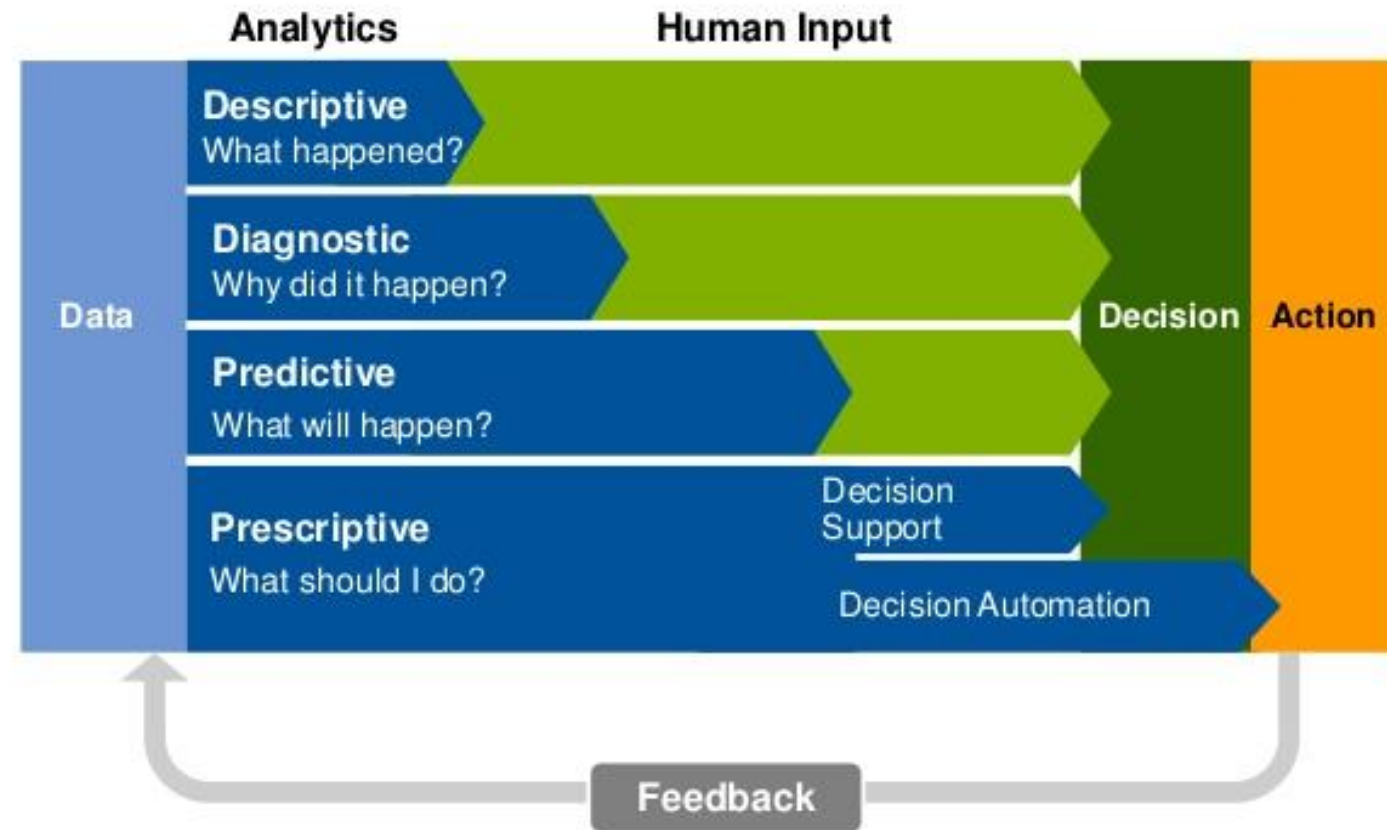
Figure 1. Magic Quadrant for Business Intelligence and Analytics Platforms



Source: Gartner (February 2017)

# Data Analytics

## The Analytics Continuum



© 2014 Gartner, Inc. and/or its affiliates. All rights reserved.

Gartner

# 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

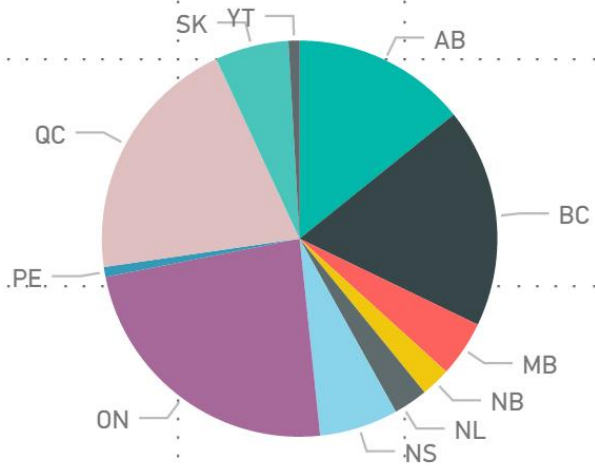
- Semi-structured
  - Job Title
  - Location
  - Company
- Unstructured
  - Job Description

## Veracity

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

# Power BI – Job Postings Demo Reports

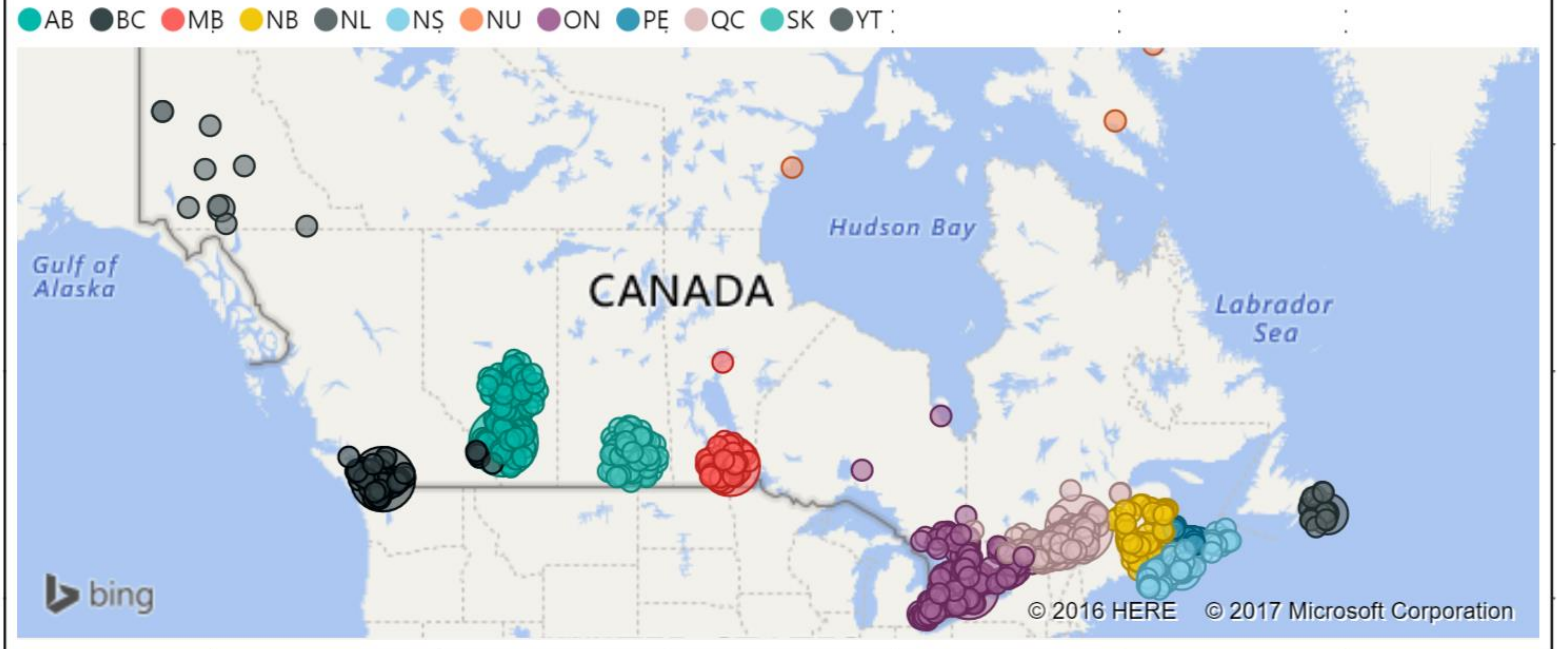
Num of Jobs by Province



21K

Num of Jobs

Jobs by Cities



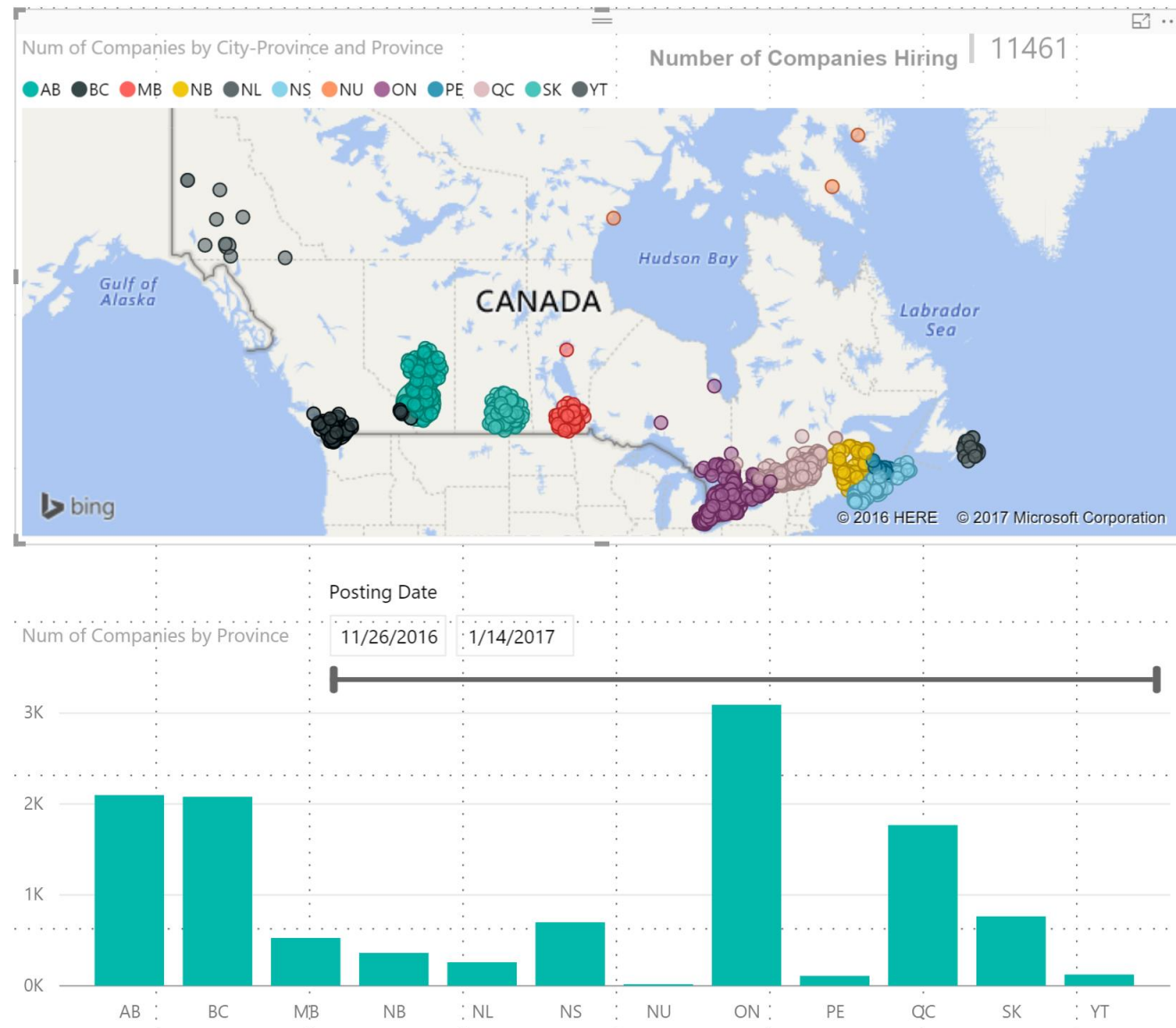
Posting Date

11/26/2016

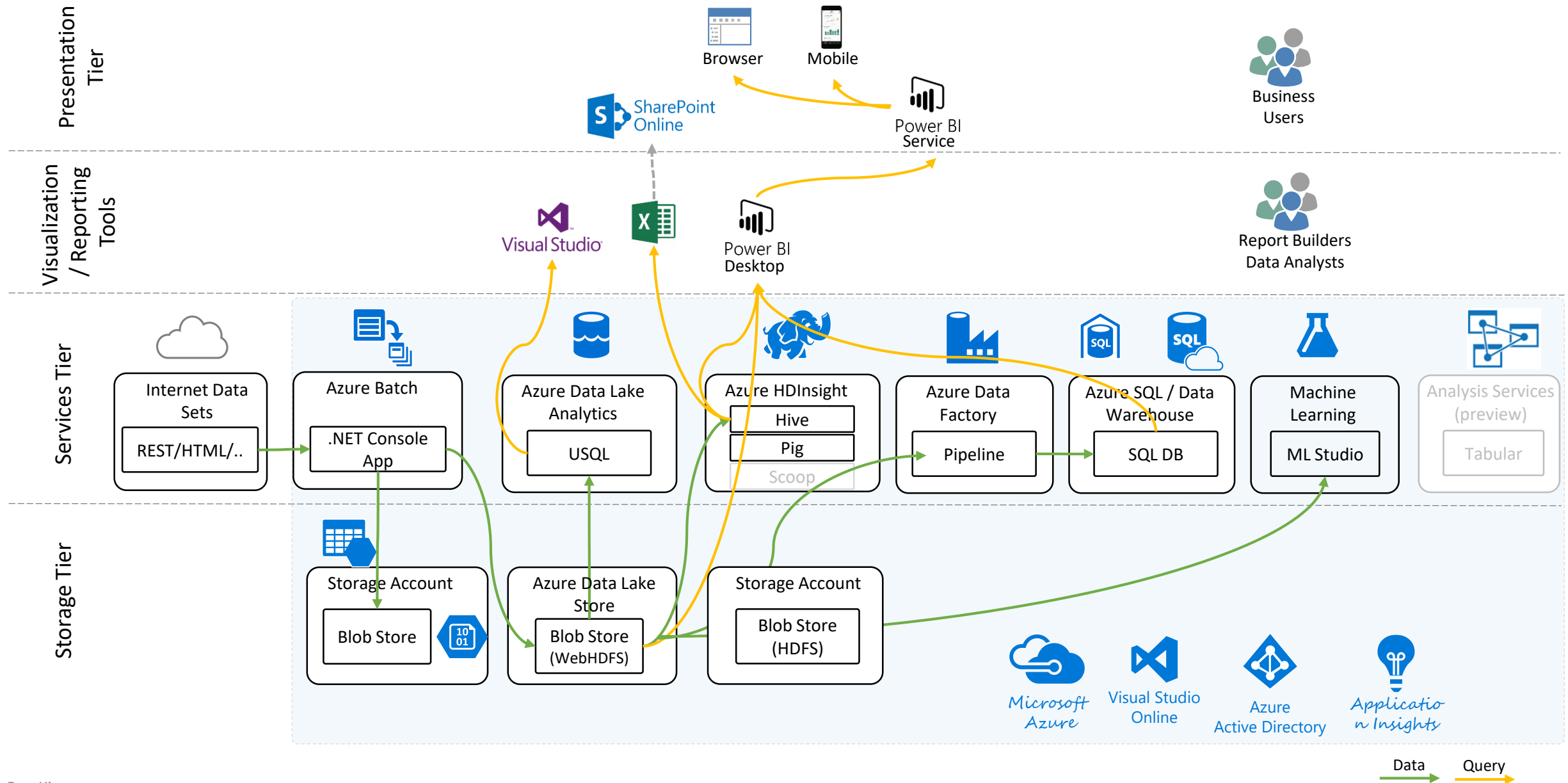
1/14/2017



# Power BI – Job Postings Demo Reports



# Job Postings Big Data Solution Architecture



## 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

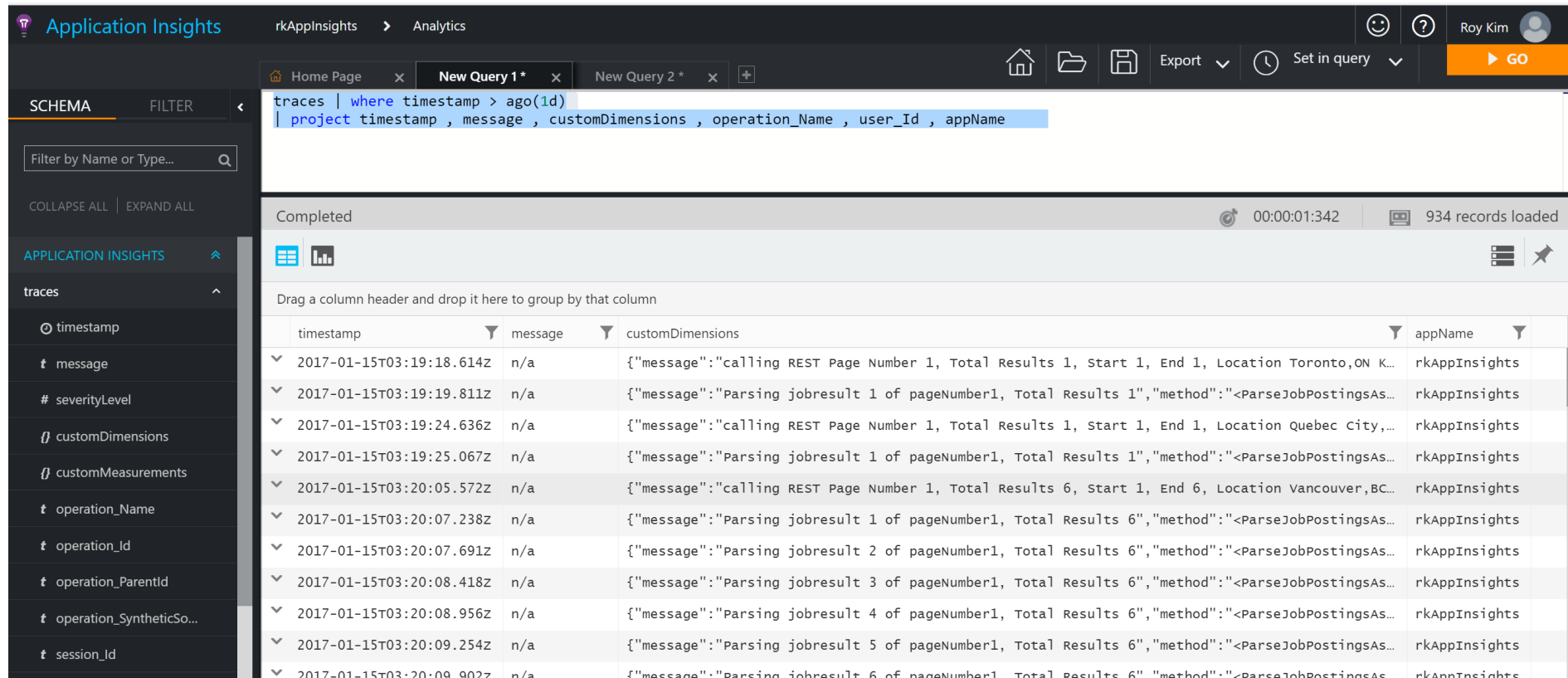
- http end points typically return JSON or XML data formats

### Html Web Page Scraping

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

# 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.



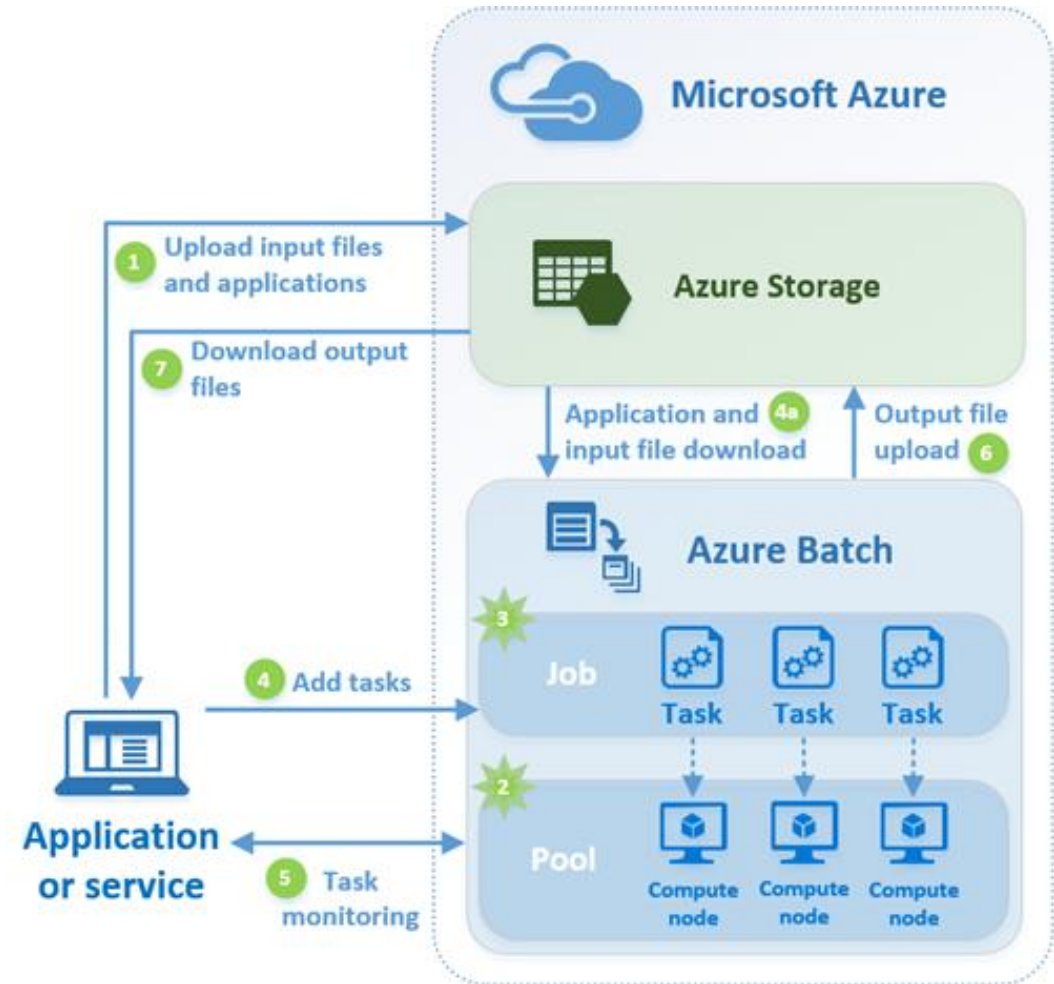
The screenshot displays the Azure Application Insights Analytics interface. The top navigation bar shows 'Application Insights' and 'rkAppInsights > Analytics'. The left sidebar contains a 'SCHEMA' section with a search filter and a list of available fields: timestamp, message, severityLevel, customDimensions, customMeasurements, operation\_Name, operation\_Id, operation\_ParentId, operation\_SyntheticSo..., and session\_Id. The main query editor shows a query: `traces | where timestamp > ago(1d) | project timestamp, message, customDimensions, operation_Name, user_Id, appName`. Below the query editor, the results are displayed in a table. The table has columns: timestamp, message, customDimensions, and appName. The results show 934 records loaded, with a completion time of 00:00:01:342. The table content is as follows:

timestamp	message	customDimensions	appName
2017-01-15T03:19:18.614Z	n/a	{"message": "calling REST Page Number 1, Total Results 1, Start 1, End 1, Location Toronto, ON K..."}	rkAppInsights
2017-01-15T03:19:19.811Z	n/a	{"message": "Parsing jobresult 1 of pageNumber1, Total Results 1", "method": "<ParseJobPostingsAs..."}	rkAppInsights
2017-01-15T03:19:24.636Z	n/a	{"message": "calling REST Page Number 1, Total Results 1, Start 1, End 1, Location Quebec City,..."}	rkAppInsights
2017-01-15T03:19:25.067Z	n/a	{"message": "Parsing jobresult 1 of pageNumber1, Total Results 1", "method": "<ParseJobPostingsAs..."}	rkAppInsights
2017-01-15T03:20:05.572Z	n/a	{"message": "calling REST Page Number 1, Total Results 6, Start 1, End 6, Location Vancouver, BC..."}	rkAppInsights
2017-01-15T03:20:07.238Z	n/a	{"message": "Parsing jobresult 1 of pageNumber1, Total Results 6", "method": "<ParseJobPostingsAs..."}	rkAppInsights
2017-01-15T03:20:07.691Z	n/a	{"message": "Parsing jobresult 2 of pageNumber1, Total Results 6", "method": "<ParseJobPostingsAs..."}	rkAppInsights
2017-01-15T03:20:08.418Z	n/a	{"message": "Parsing jobresult 3 of pageNumber1, Total Results 6", "method": "<ParseJobPostingsAs..."}	rkAppInsights
2017-01-15T03:20:08.956Z	n/a	{"message": "Parsing jobresult 4 of pageNumber1, Total Results 6", "method": "<ParseJobPostingsAs..."}	rkAppInsights
2017-01-15T03:20:09.254Z	n/a	{"message": "Parsing jobresult 5 of pageNumber1, Total Results 6", "method": "<ParseJobPostingsAs..."}	rkAppInsights
2017-01-15T03:20:09.902Z	n/a	{"message": "Parsing jobresult 6 of pageNumber1, Total Results 6", "method": "<ParseJobPostingsAs..."}	rkAppInsights



# Azure Batch

- 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



# 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

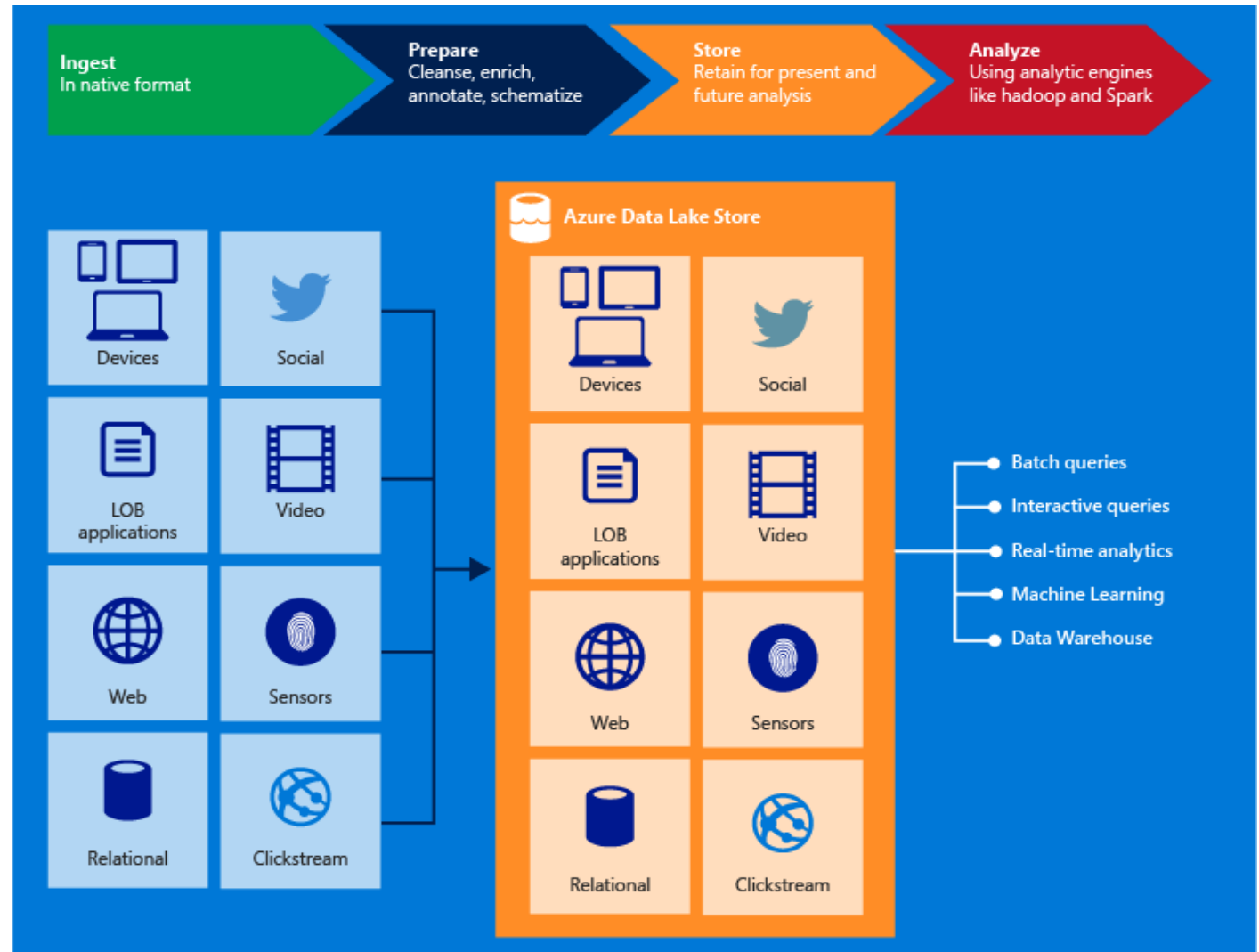
# Azure Data Lake Store

- 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

# Azure Data Lake Store

## Use Cases

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



# 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.

# Azure Data Lake Analytics – Demo Implementation

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

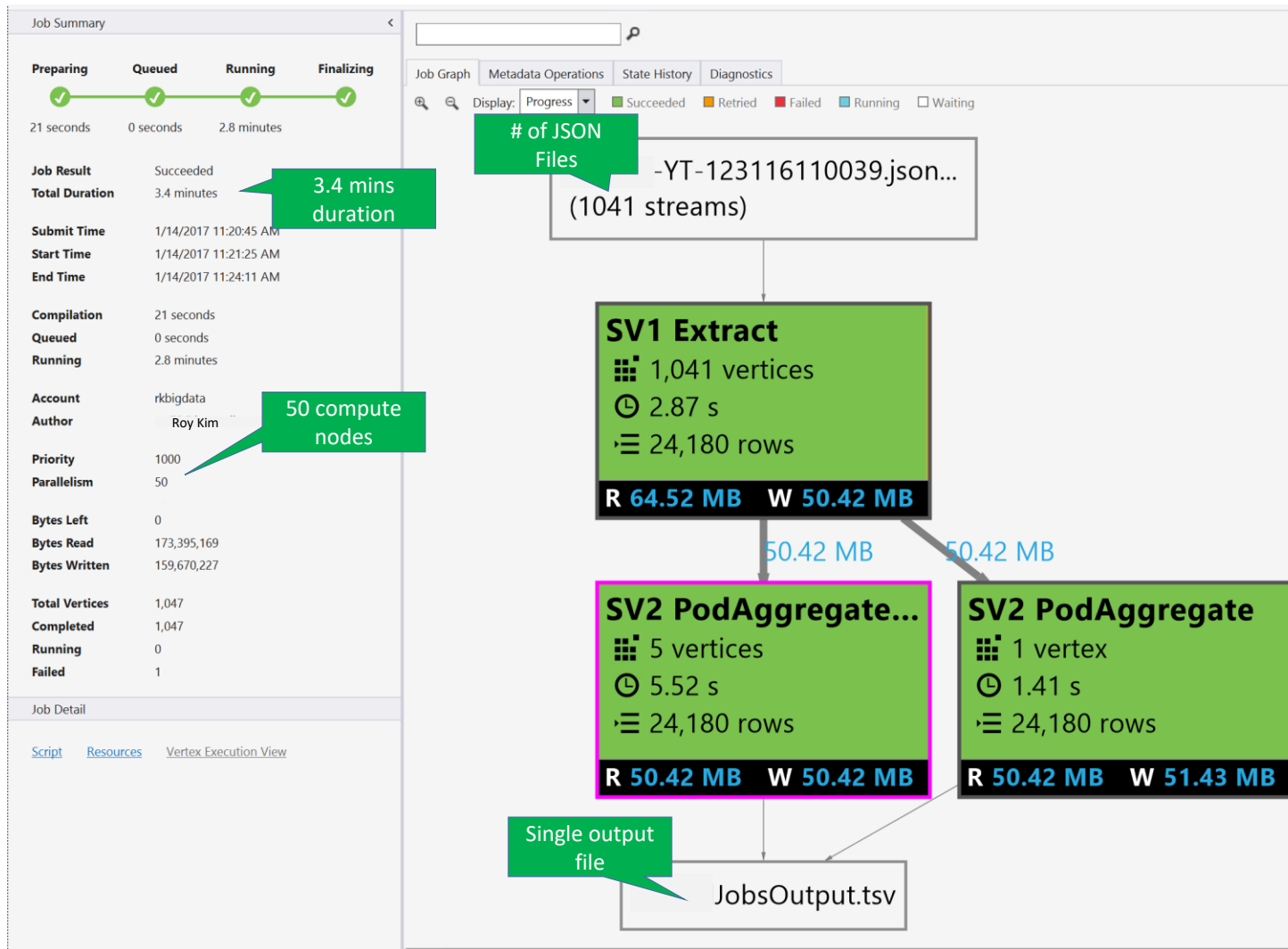
```
REFERENCE ASSEMBLY [Newtonsoft.Json];
REFERENCE ASSEMBLY [Microsoft.Analytics.Samples.Formats];

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,
        jobType string
FROM @inputfile
USING new Microsoft.Analytics.Samples.Formats.Json.JsonExtractor("Results[*]");

OUTPUT @jobPostingsSchema
TO "/jobpostings/outputtsv/v3/JobsOutput.tsv"
USING Outputters.Tsv();
```

# Azure Data Lake Analytics – Demo Implementation



## Azure HDInsight

- 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.



# Azure HD Insight – Demo Implementation

## 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.

## Azure SQL Database

- 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.

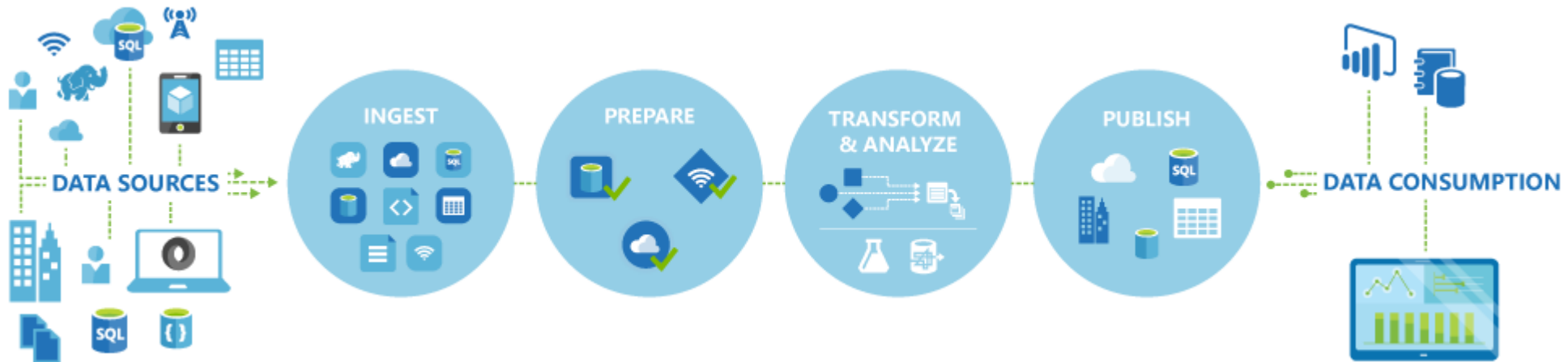
# Azure SQL Database – Demo Implementation

- 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

Year	MonthName	Week... ▲	DayName	AB	BC	NS	ON	QC	SK	Total
2016	December	1	Friday	67	15	42	62	76	15	277
			Saturday	21	6	6	25	27	5	90
			Thursday	87	21	41	72	52	16	289
			<b>Total</b>	<b>175</b>	<b>42</b>	<b>89</b>	<b>159</b>	<b>155</b>	<b>36</b>	<b>656</b>
		2	Friday	43	23	21	18	80	46	231
			Monday	58	12	24	39	74	19	226
			Saturday	42	10	7	11	34	17	121
			Sunday	7	8	5	6	43	2	71
			Thursday	95	46	31	35	92	8	307
			Tuesday	83	361	21	100	79	14	658
			Wednesday	153	27	44	117	67	2	410
			<b>Total</b>	<b>481</b>	<b>487</b>	<b>153</b>	<b>326</b>	<b>469</b>	<b>108</b>	<b>2024</b>
		3	Friday	87	44	31	35	118	45	360
			Monday	127	6	79	283	97	31	623
			Saturday	80	24	17	14	48	8	191
			Sunday	9	6	3	8	27	7	60
			Thursday	86	29	47	28	130	25	345
			Tuesday	81	48	36	31	112	33	341
			Wednesday	27	53	33	27	113	29	282
			<b>Total</b>	<b>497</b>	<b>210</b>	<b>246</b>	<b>426</b>	<b>645</b>	<b>178</b>	<b>2202</b>

# 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.)

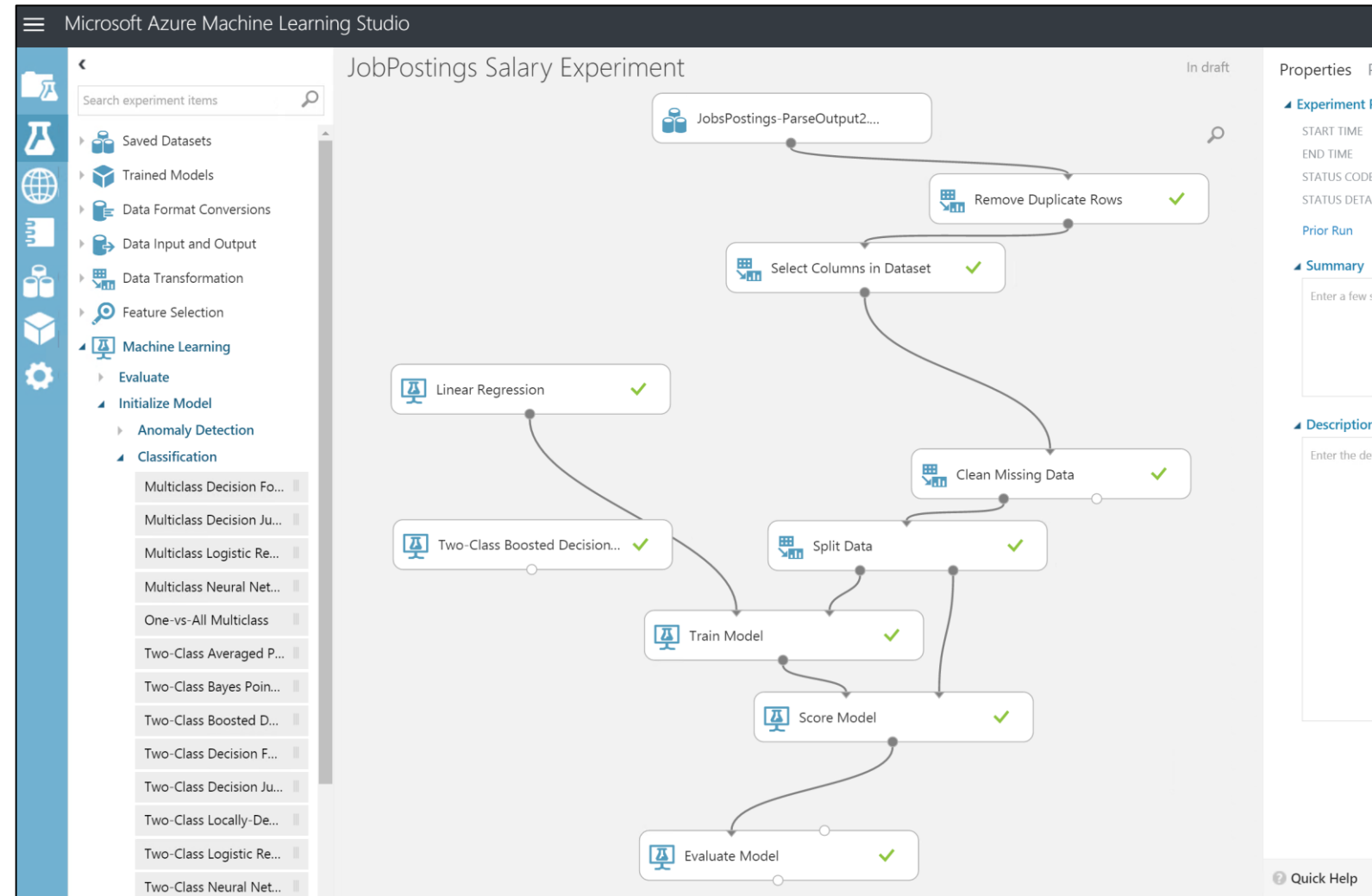


# Azure Data Factory

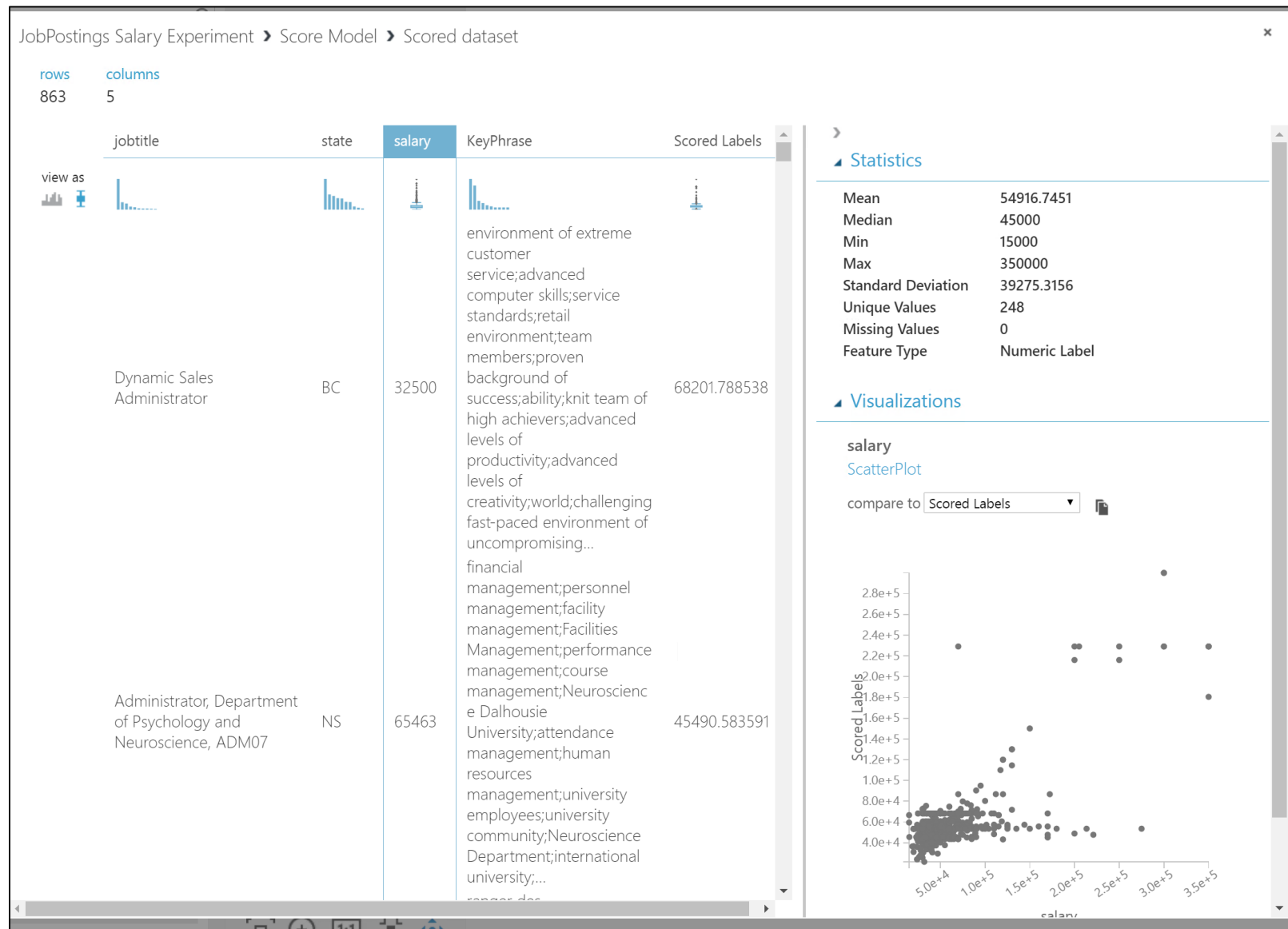
Category	Data store	Supported as a source	Supported as a sink
Azure	<a href="#">Azure Blob storage</a>	✓	✓
	<a href="#">Azure Data Lake Store</a>	✓	✓
	<a href="#">Azure SQL Database</a>	✓	✓
	<a href="#">Azure SQL Data Warehouse</a>	✓	✓
	<a href="#">Azure Table storage</a>	✓	✓
	<a href="#">Azure DocumentDB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

# Azure Machine Learning – Demo Implementation

Predicting Salary for a given set of parameters such as job title and location



# Azure Machine Learning – Demo Implementation



# Power BI App Service

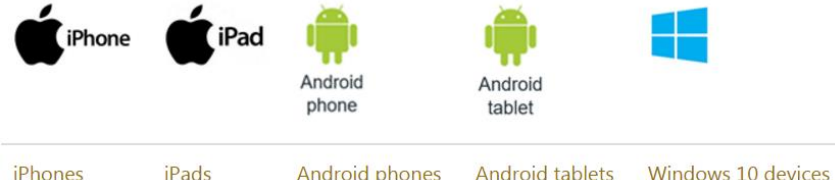
The main features of your Power BI service UI:

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



## 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

# Mobile App IOS Key Features & Demo

- Navigation
- Dashboards and Reports
- Responsive design
- Visualization interaction
- Sharing
- Annotations
- Q&A
- Alerts
- Favourites



Annotations

## Closing Remarks

- 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.

## Q&A



@RoyKimYYZ

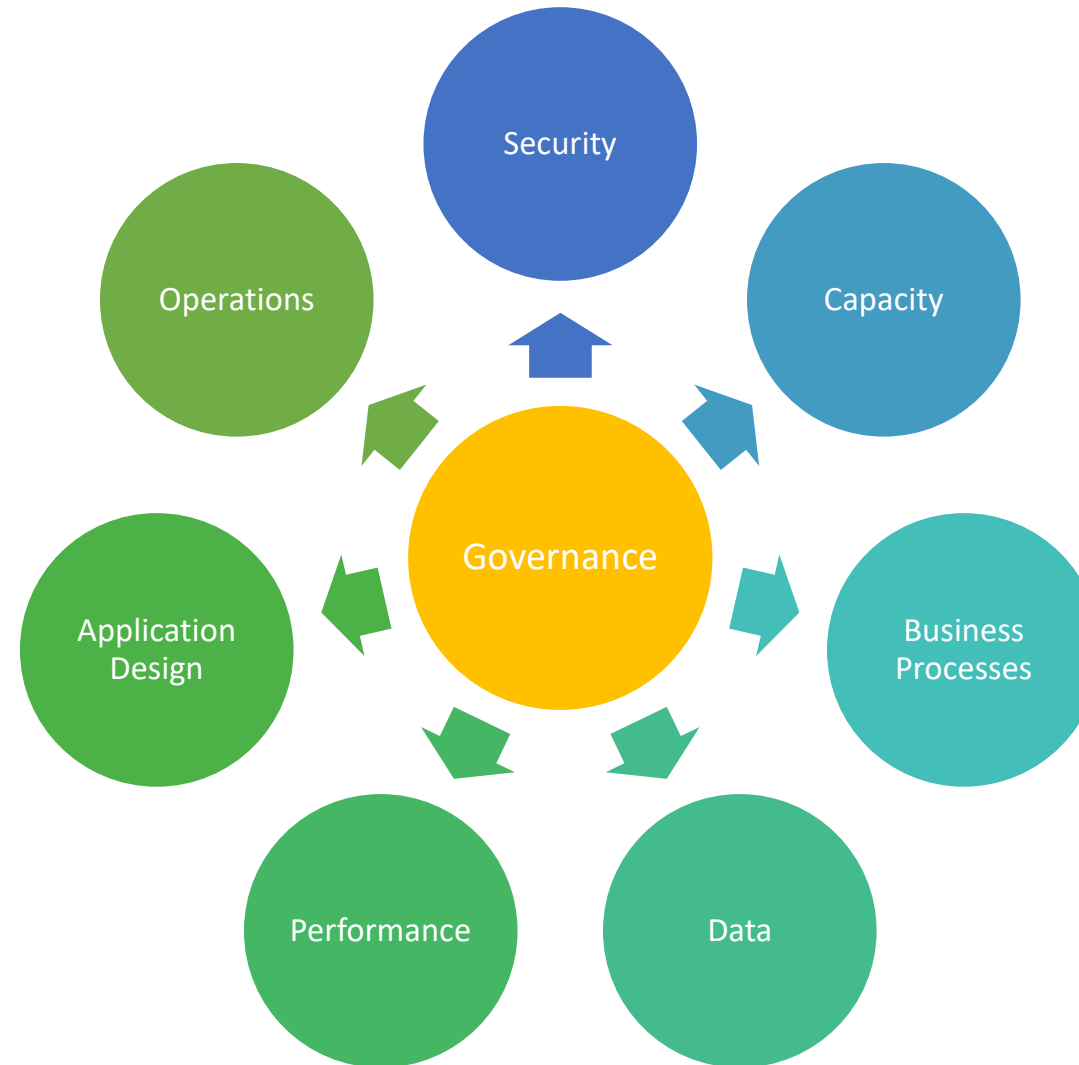


[roykimtoronto@gmail.com](mailto:roykimtoronto@gmail.com)

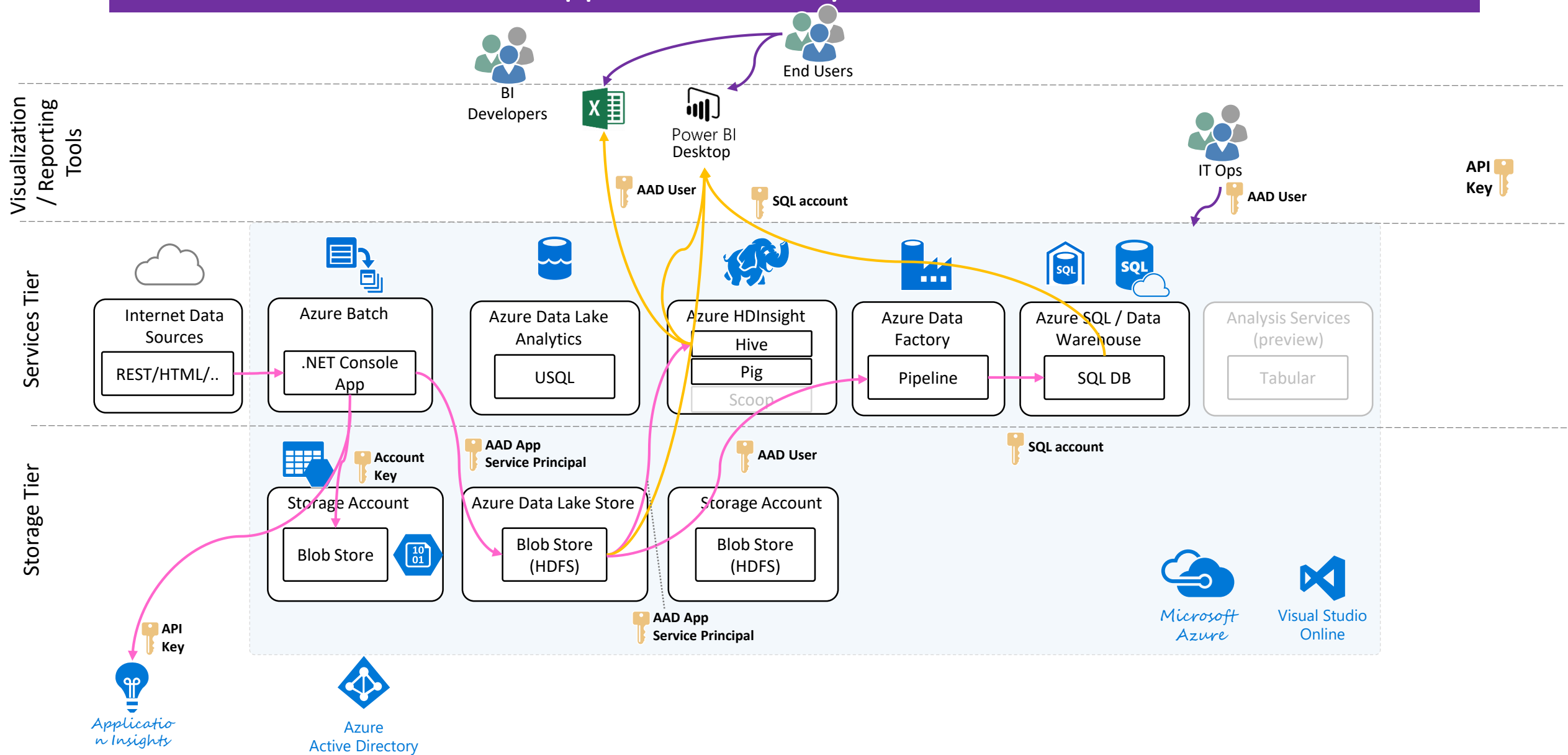


roykim.ca

## Appendix - Architecture



# Appendix - Security Architecture



# Appendix - Data Processing & Formats

