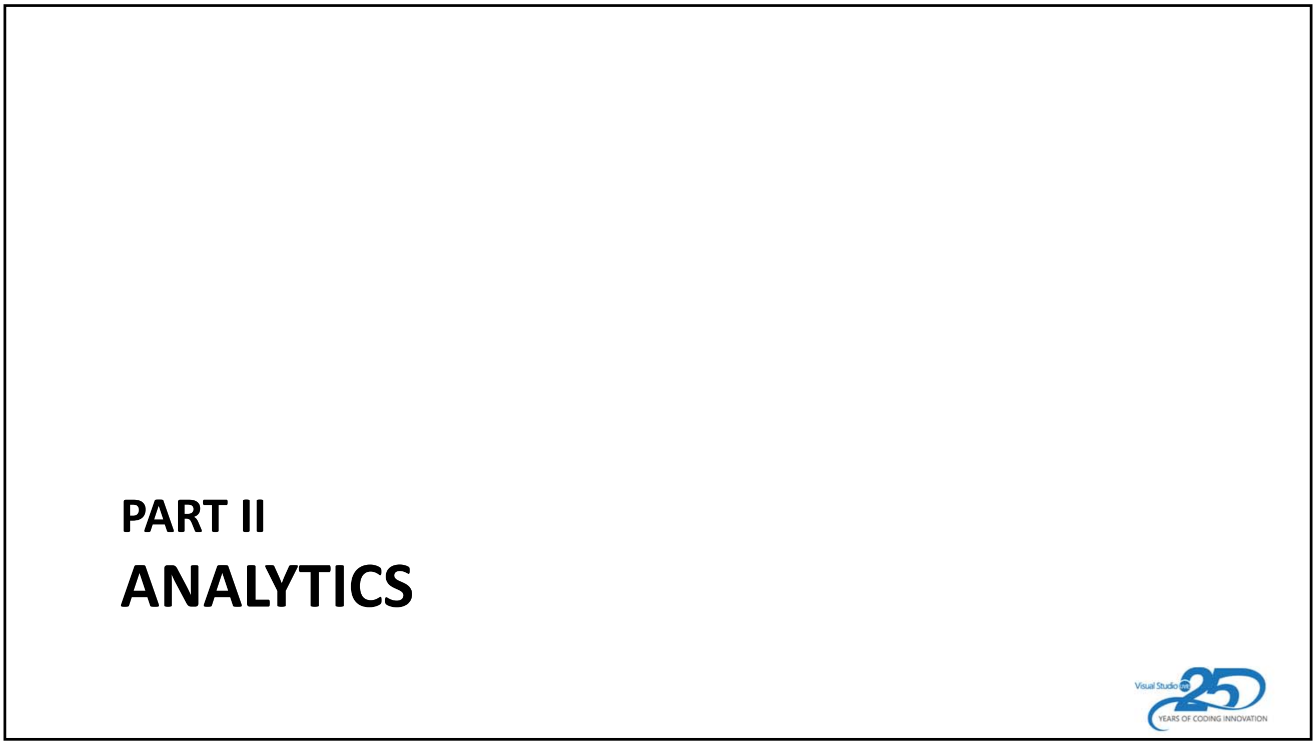
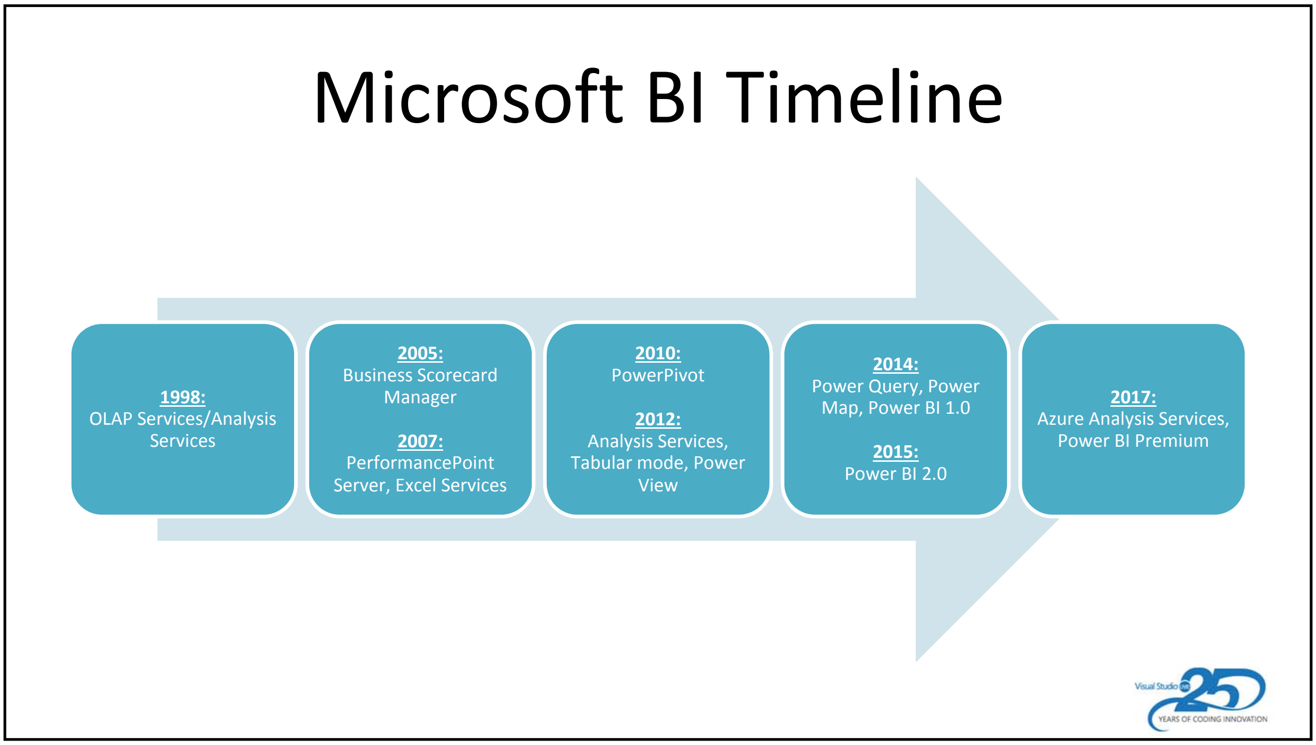


PART II

ANALYTICS



Microsoft BI Timeline

1998:
OLAP Services/Analysis Services

2005:
Business Scorecard Manager

2007:
PerformancePoint Server, Excel Services

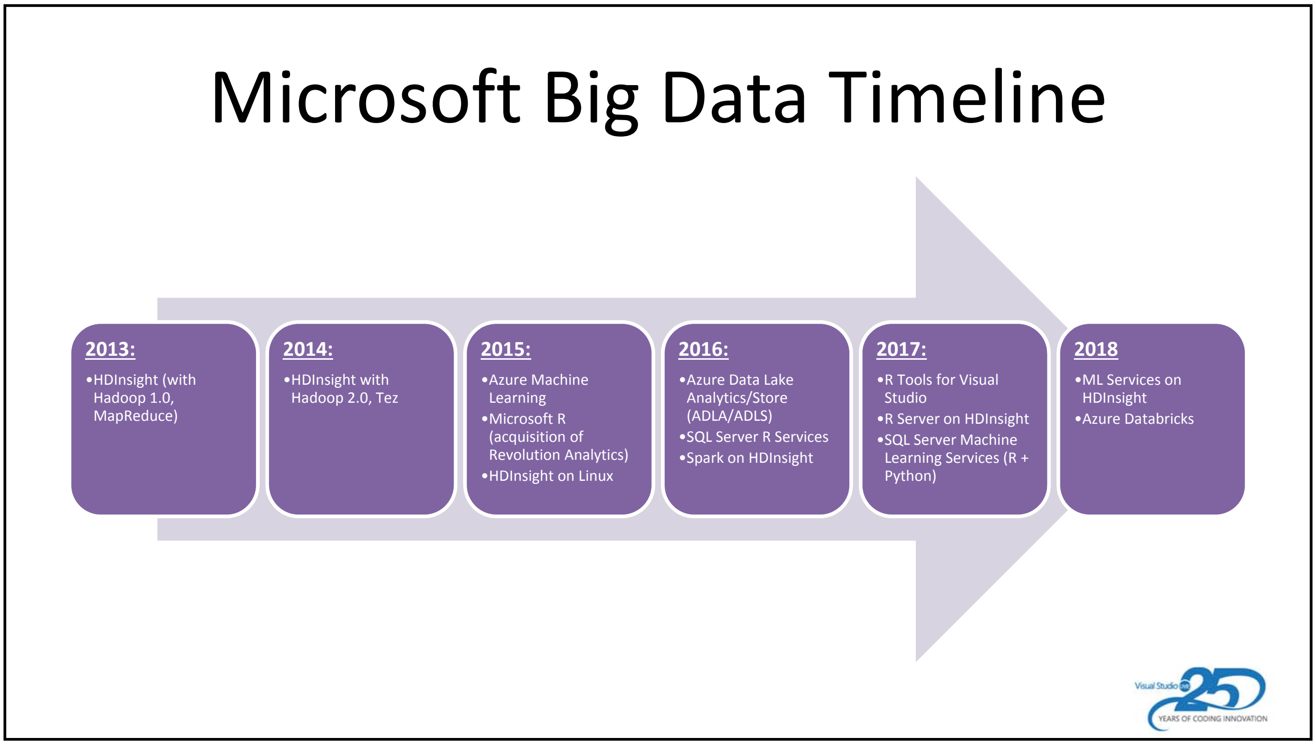
2010:
PowerPivot

2012:
Analysis Services, Tabular mode, Power View

2014:
Power Query, Power Map, Power BI 1.0

2015:
Power BI 2.0

2017:
Azure Analysis Services, Power BI Premium



Microsoft Big Data Timeline

2013:
- HDInsight (with Hadoop 1.0, MapReduce)

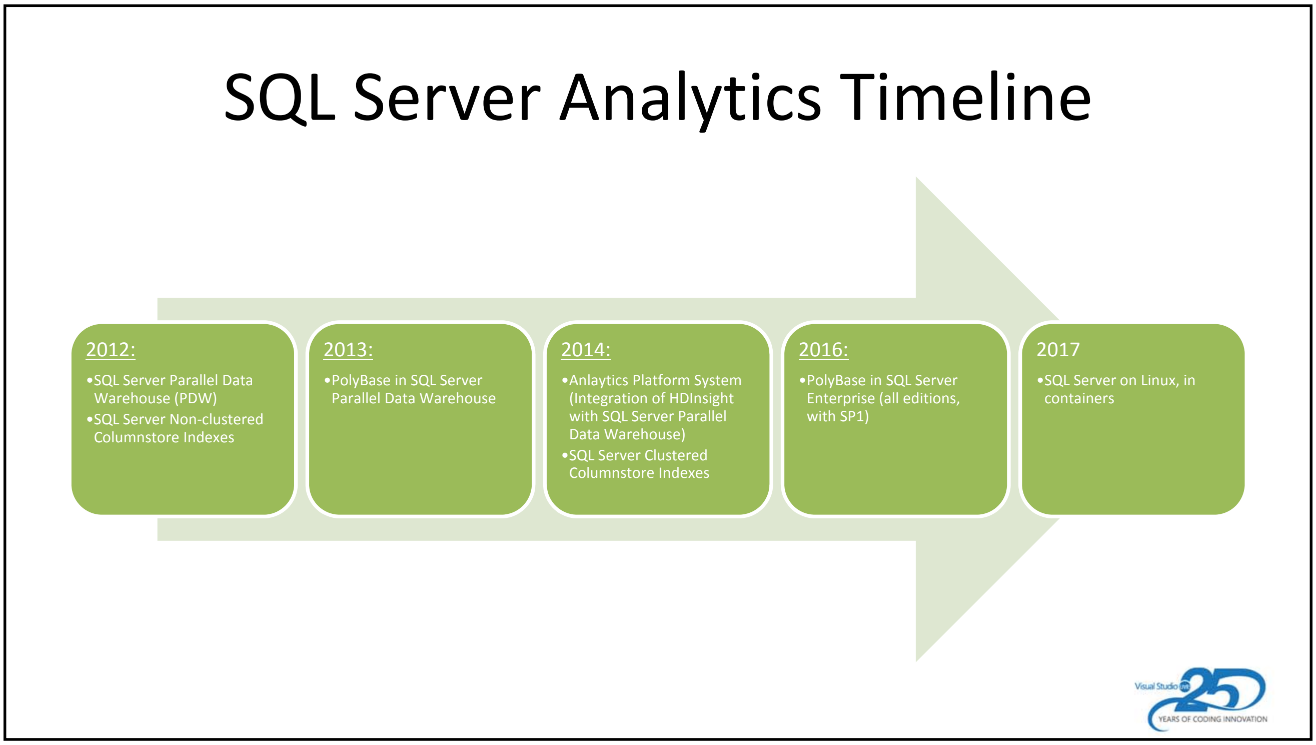
2014:
- HDInsight with Hadoop 2.0, Tez

2015:
- Azure Machine Learning
- Microsoft R (acquisition of Revolution Analytics)
- HDInsight on Linux

2016:
- Azure Data Lake Analytics/Store (ADLA/ADLS)
- SQL Server R Services
- Spark on HDInsight

2017:
- R Tools for Visual Studio
- R Server on HDInsight
- SQL Server Machine Learning Services (R + Python)

2018
- ML Services on HDInsight
- Azure Databricks



SQL Server Analytics Timeline

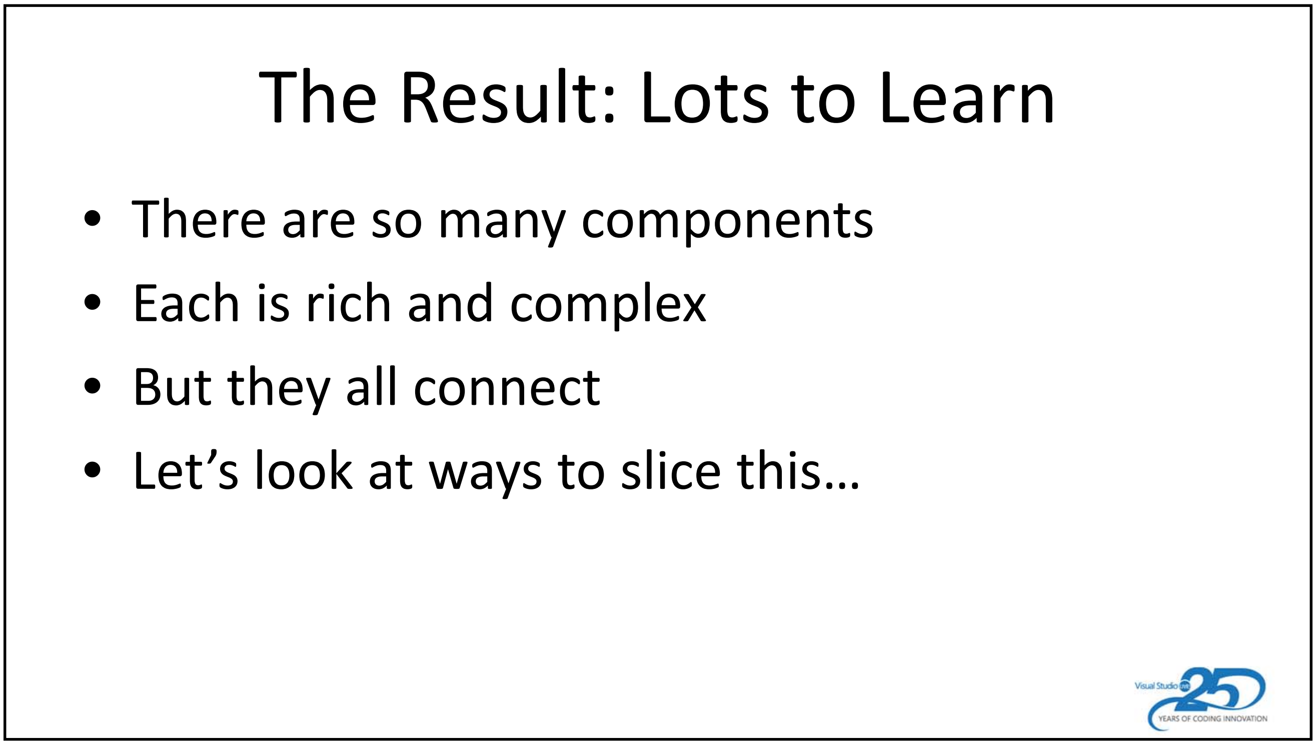
2012:
- SQL Server Parallel Data Warehouse (PDW)
- SQL Server Non-clustered Columnstore Indexes

2013:
- PolyBase in SQL Server Parallel Data Warehouse

2014:
- Anlaytics Platform System (Integration of HDInsight with SQL Server Parallel Data Warehouse)
- SQL Server Clustered Columnstore Indexes

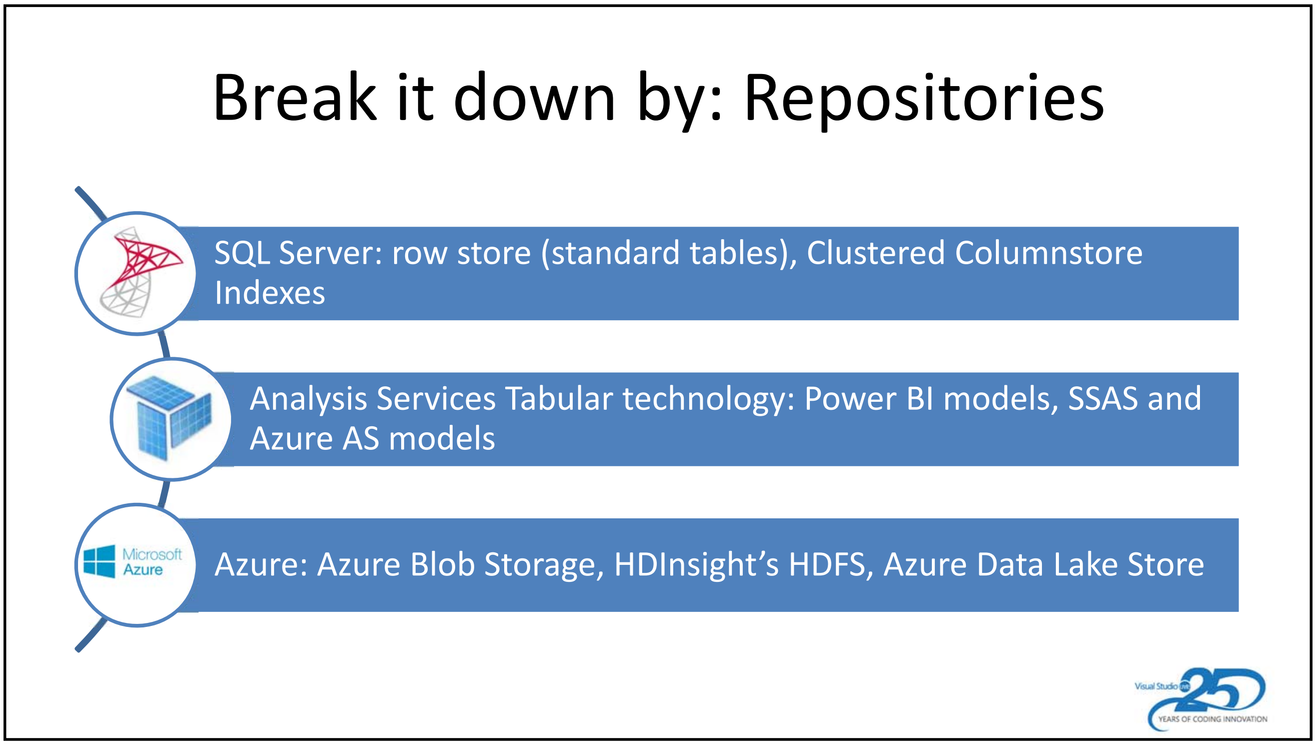
2016:
- PolyBase in SQL Server Enterprise (all editions, with SP1)

2017
- SQL Server on Linux, in containers



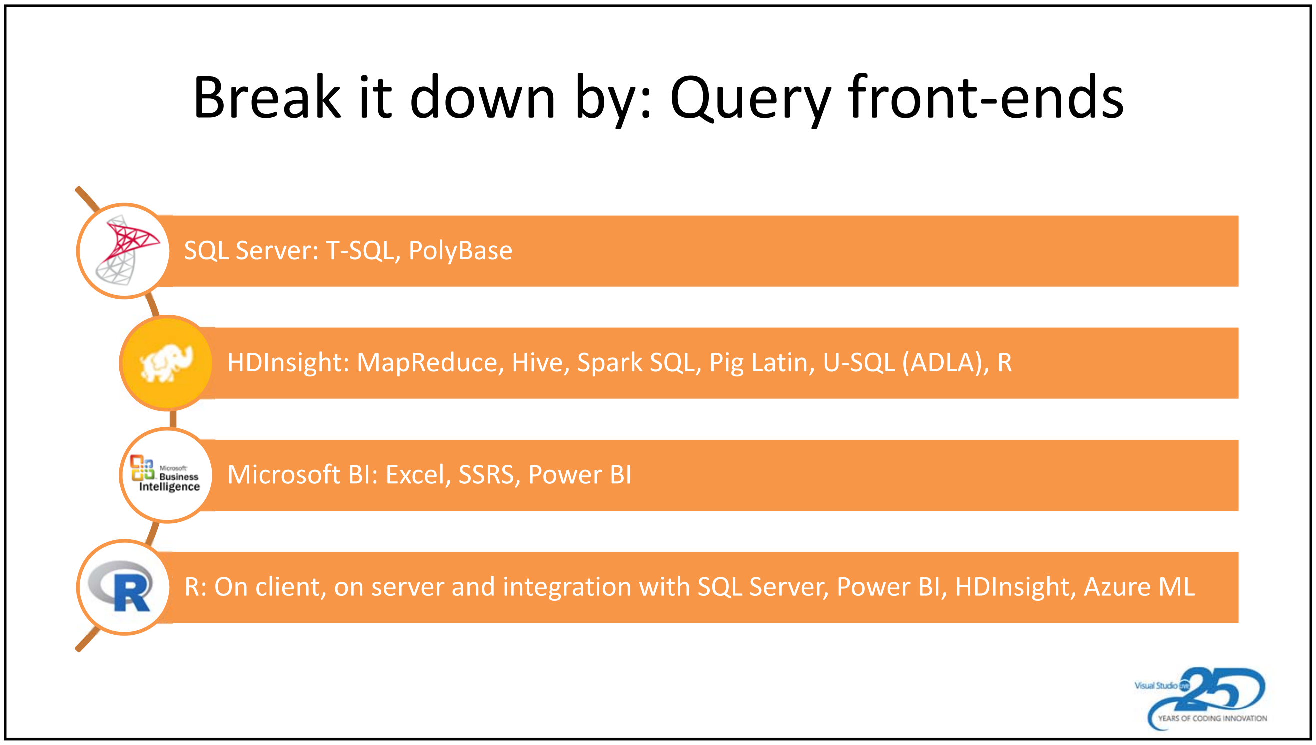
The Result: Lots to Learn

- There are so many components
- Each is rich and complex
- But they all connect
- Let’s look at ways to slice this…



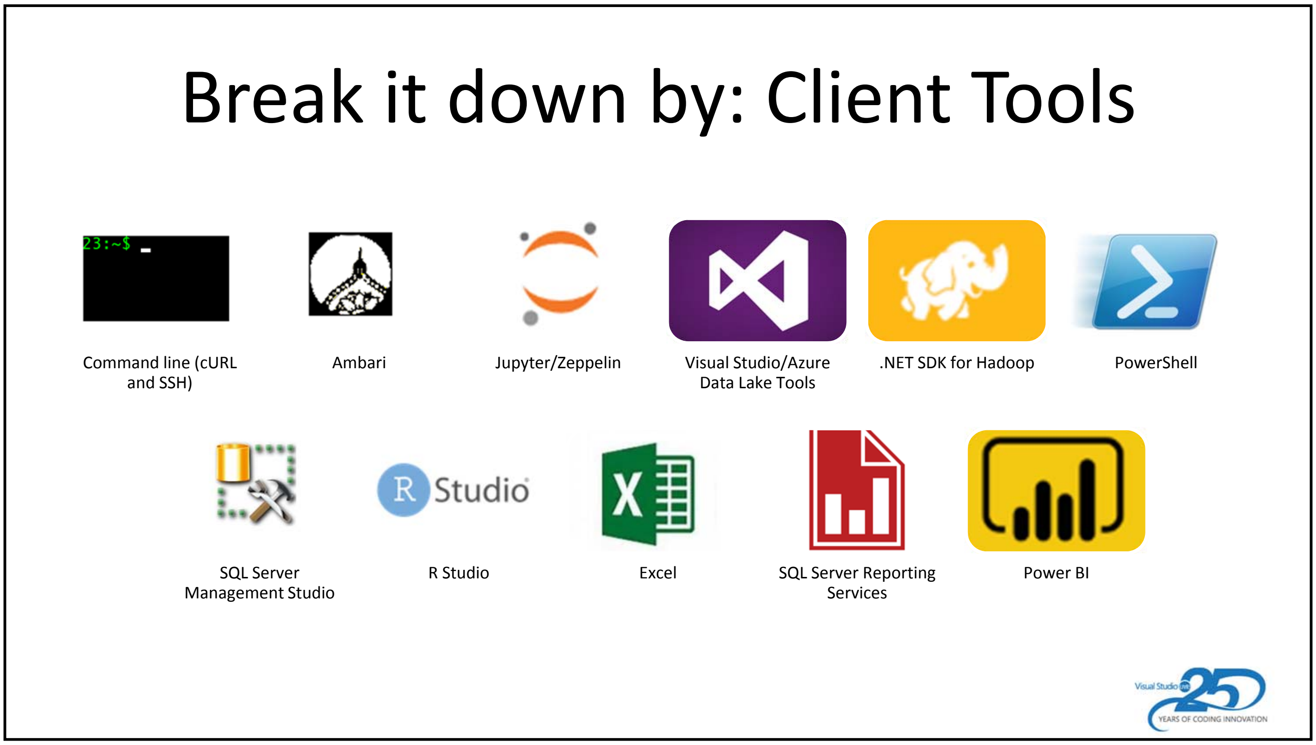
Break it down by: Repositories

- SQL Server: row store (standard tables), Clustered Columnstore Indexes
- Analysis Services Tabular technology: Power BI models, SSAS and Azure AS models
- Azure: Azure Blob Storage, HDInsight’s HDFS, Azure Data Lake Store



Break it down by: Query front-ends

- SQL Server: T-SQL, PolyBase
- HDInsight: MapReduce, Hive, Spark SQL, Pig Latin, U-SQL (ADLA), R
- Microsoft BI: Excel, SSRS, Power BI
- R: On client, on server and integration with SQL Server, Power BI, HDInsight, Azure ML



Break it down by: Client Tools

Command line (cURL and SSH)

Ambari

Jupyter/Zeppelin

Visual Studio/Azure Data Lake Tools

.NET SDK for Hadoop

PowerShell

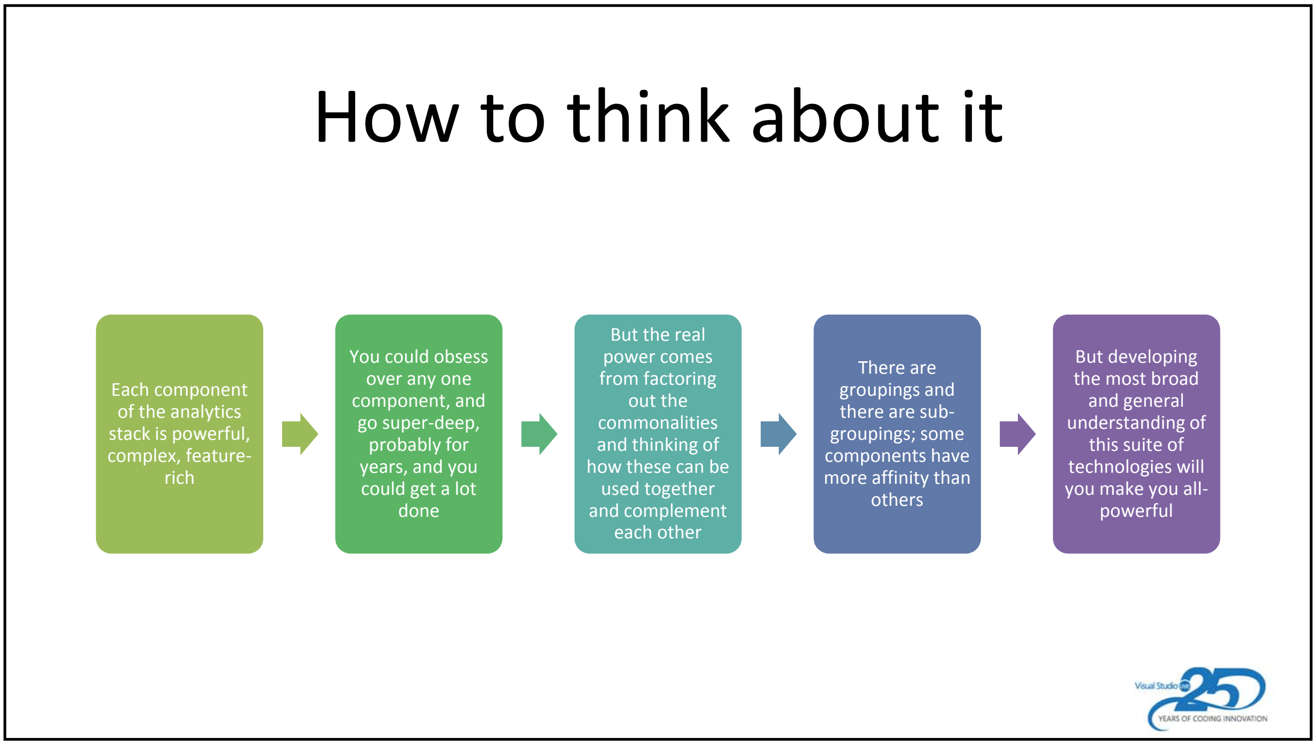
SQL Server Management Studio

R Studio

Excel

SQL Server Reporting Services

Power BI



How to think about it

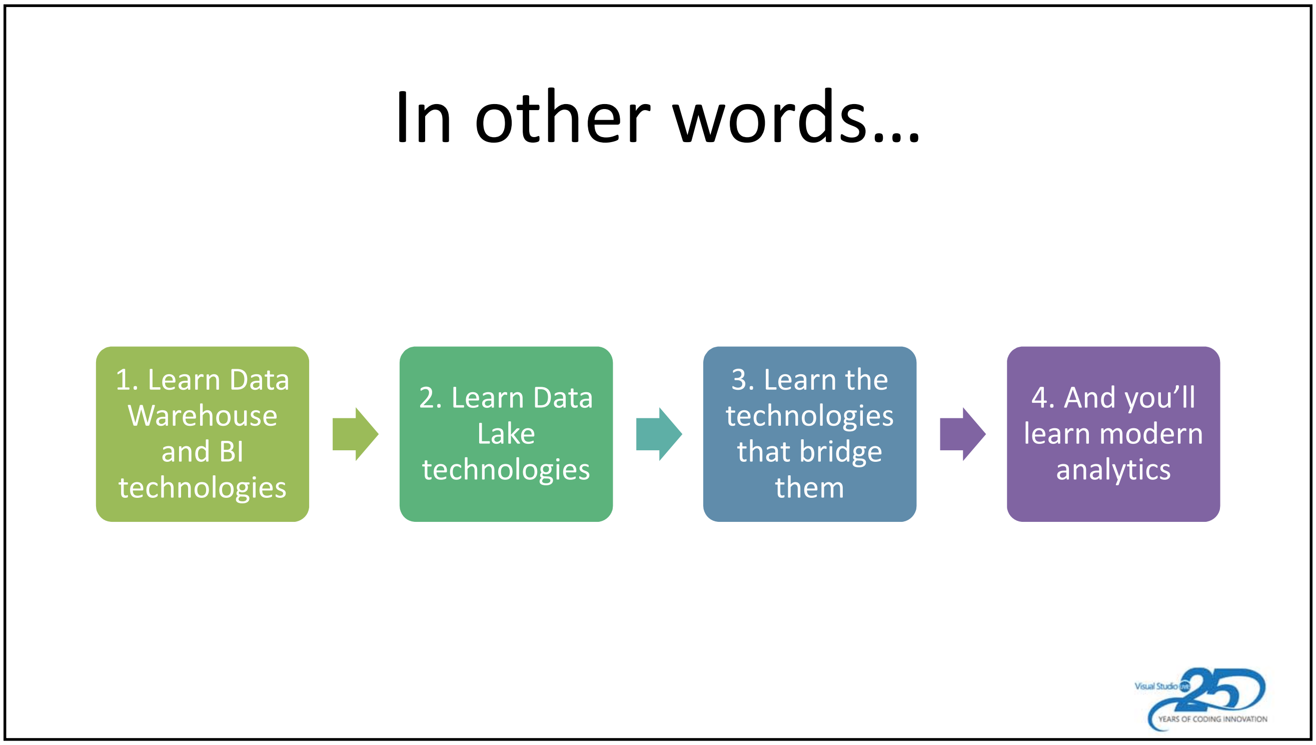
Each component of the analytics stack is powerful, complex, feature-rich

You could obsess over any one component, and go super-deep, probably for years, and you could get a lot done

But the real power comes from factoring out the commonalities and thinking of how these can be used together and complement each other

There are groupings and there are sub-groupings; some components have more affinity than others

But developing the most broad and general understanding of this suite of technologies will you make you all-powerful



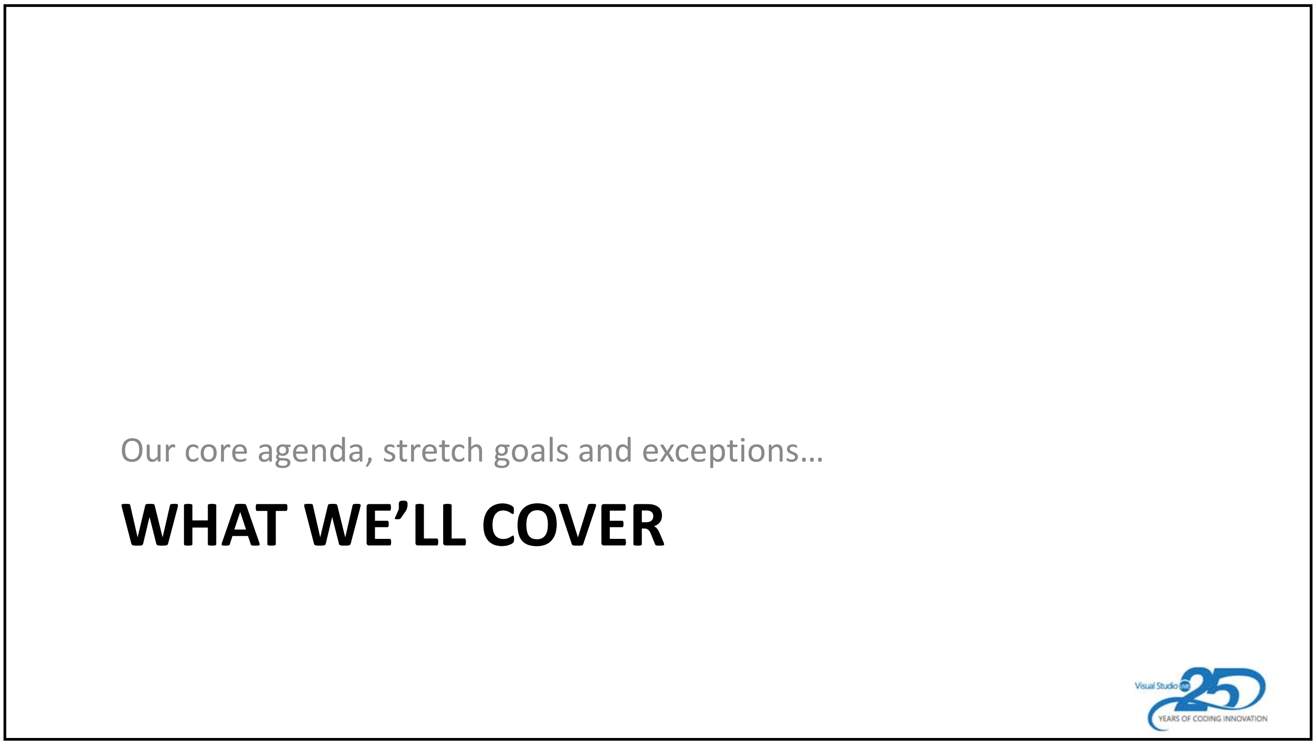
In other words…

1. Learn Data Warehouse and BI technologies

2. Learn Data Lake technologies

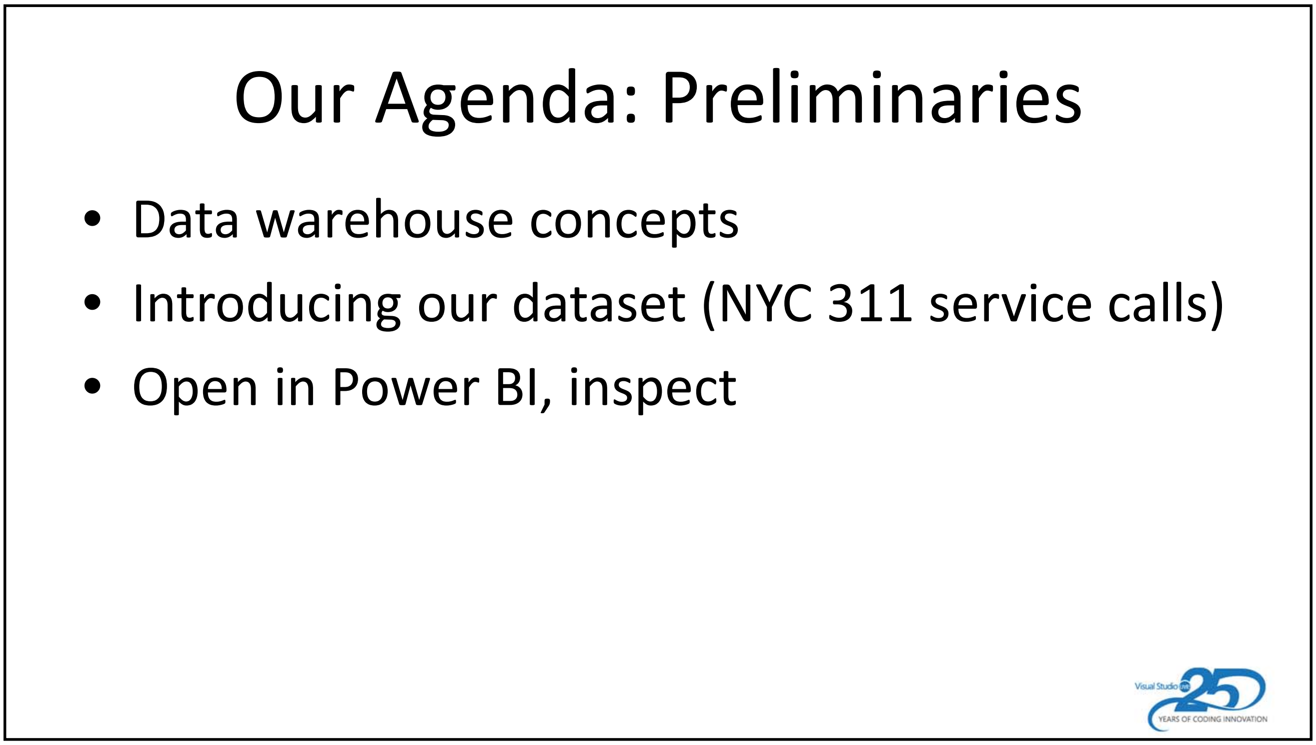
3. Learn the technologies that bridge them

4. And you'll learn modern analytics



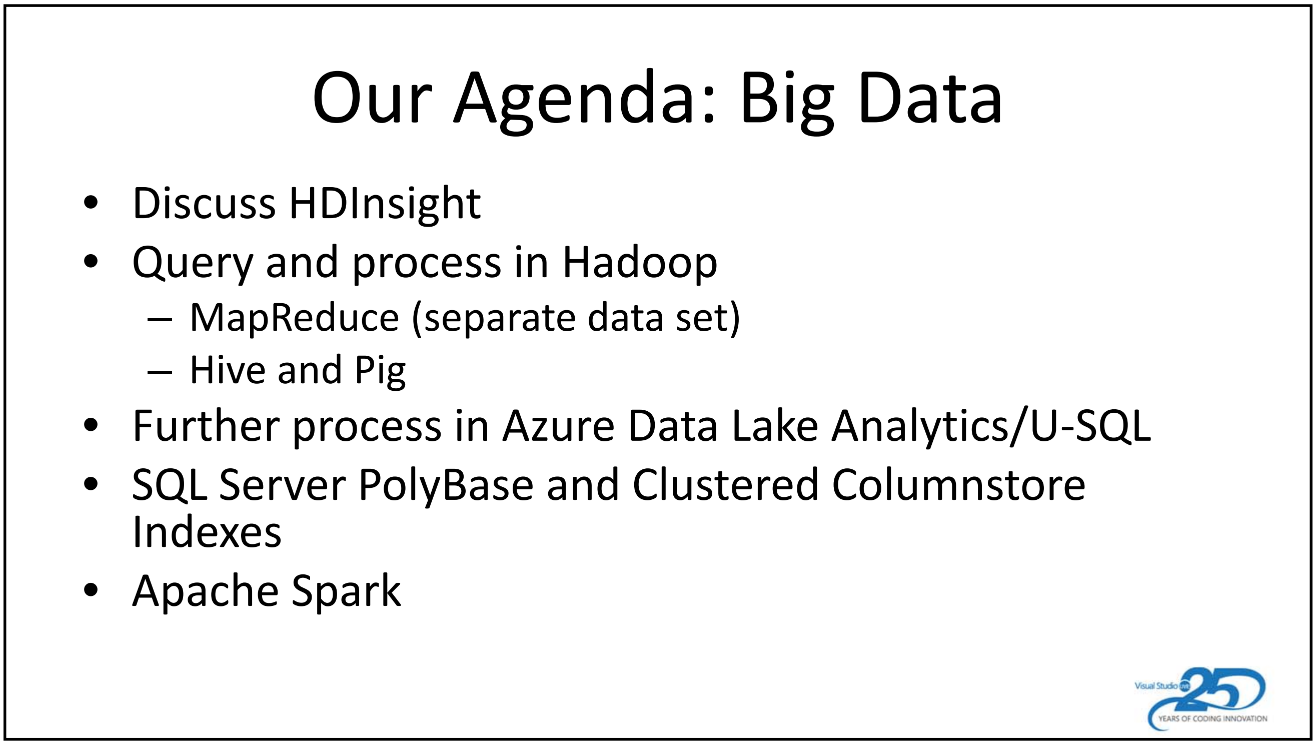
Our core agenda, stretch goals and exceptions…

WHAT WE'LL COVER



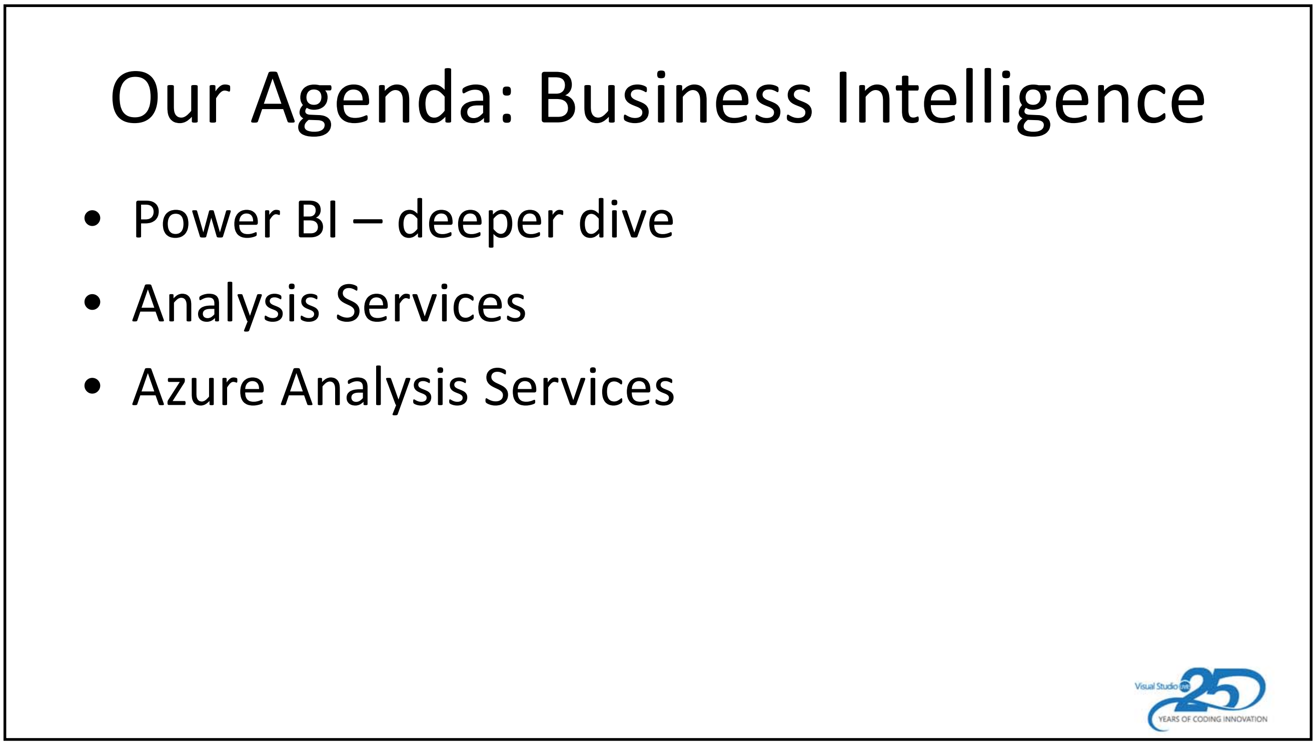
Our Agenda: Preliminaries

- Data warehouse concepts
- Introducing our dataset (NYC 311 service calls)
- Open in Power BI, inspect



Our Agenda: Big Data

- Discuss HDInsight
- Query and process in Hadoop
  - MapReduce (separate data set)
  - Hive and Pig
- Further process in Azure Data Lake Analytics/U-SQL
- SQL Server PolyBase and Clustered Columnstore Indexes
- Apache Spark



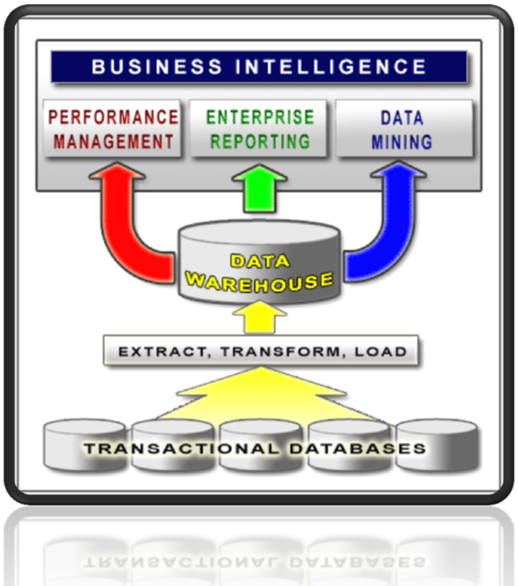
Our Agenda: Business Intelligence

- Power BI – deeper dive
- Analysis Services
- Azure Analysis Services

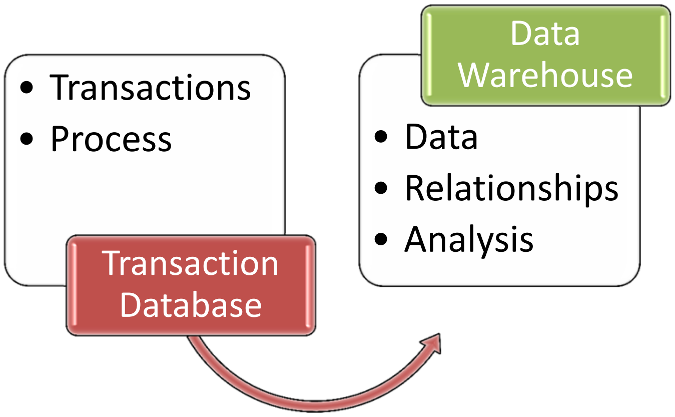
Dimensional Analysis and MPP

DATA WAREHOUSE CONCEPTS

Business Intelligence

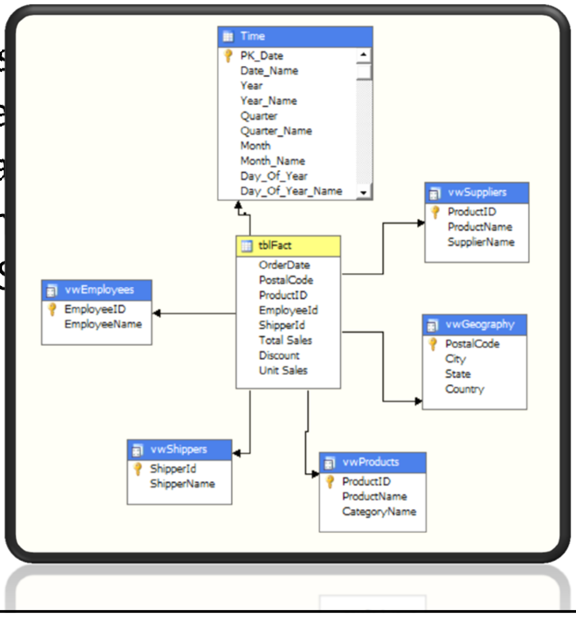


Preparing For Business Intelligence



Dimensional Model

- Meas
- Dime
- Hiera
- Grair
- Star S


Time
PK_Date
Date_Name
Year
Year_Name
Quarter
Quarter_Name
Month
Month_Name
Day_Of_Year
Day_Of_Year_Name

tblFact
OrderDate
PostalCode
ProductID
EmployeeId
ShipperId
Total Sales
Discount
Unit Sales

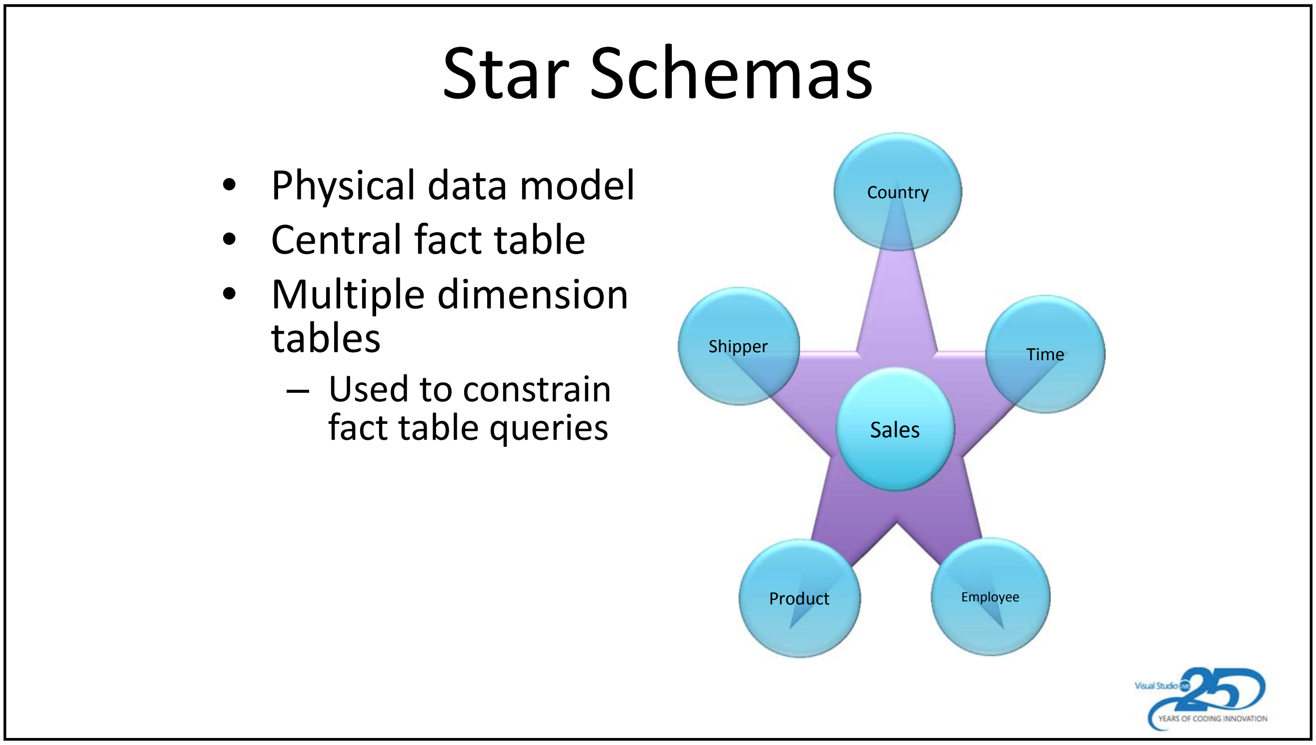
vwSuppliers
ProductID
ProductName
SupplierName

vwEmployees
EmployeeID
EmployeeName

vwGeography
PostalCode
City
State
Country

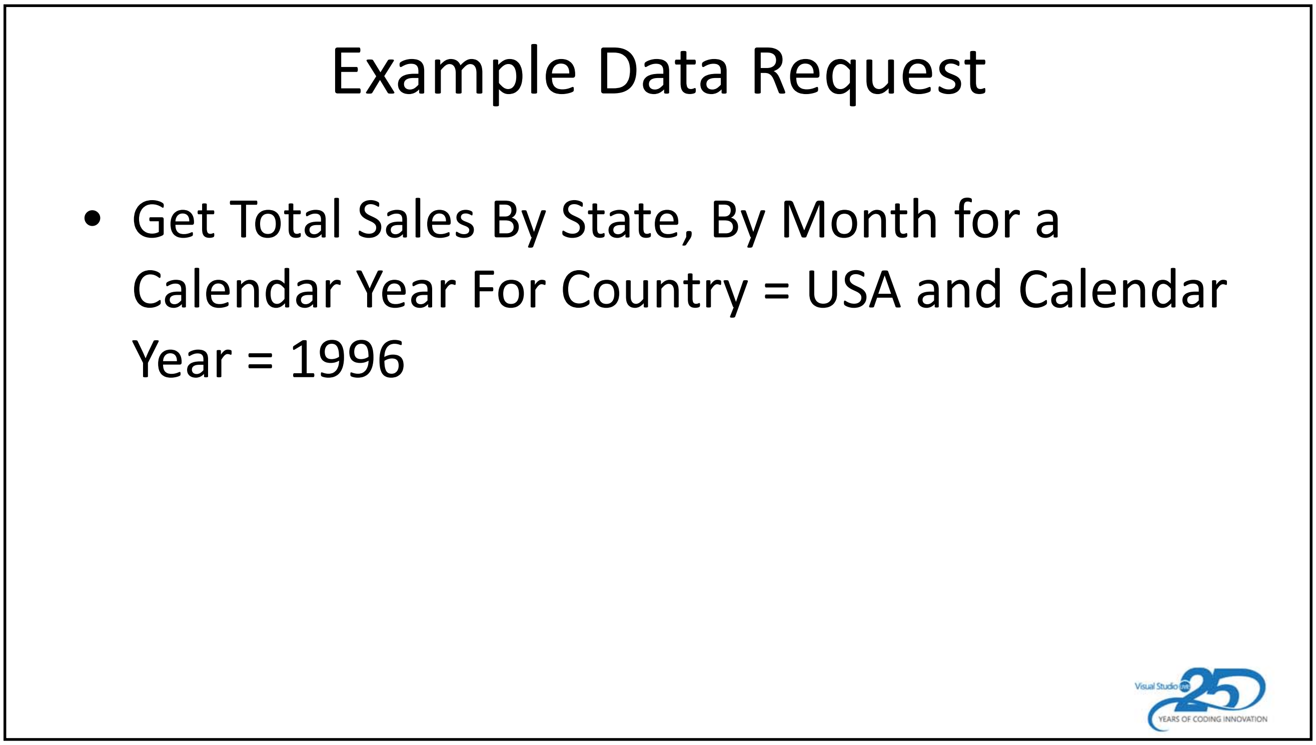
vwShippers
ShipperId
ShipperName

vwProducts
ProductID
ProductName
CategoryName



Star Schemas

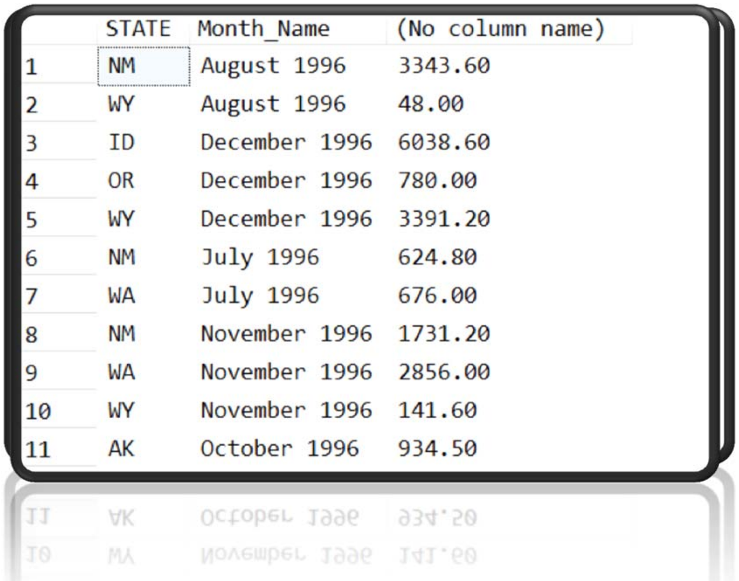
- Physical data model
- Central fact table
- Multiple dimension tables
  - Used to constrain fact table queries



Example Data Request

- Get Total Sales By State, By Month for a Calendar Year For Country = USA and Calendar Year = 1996

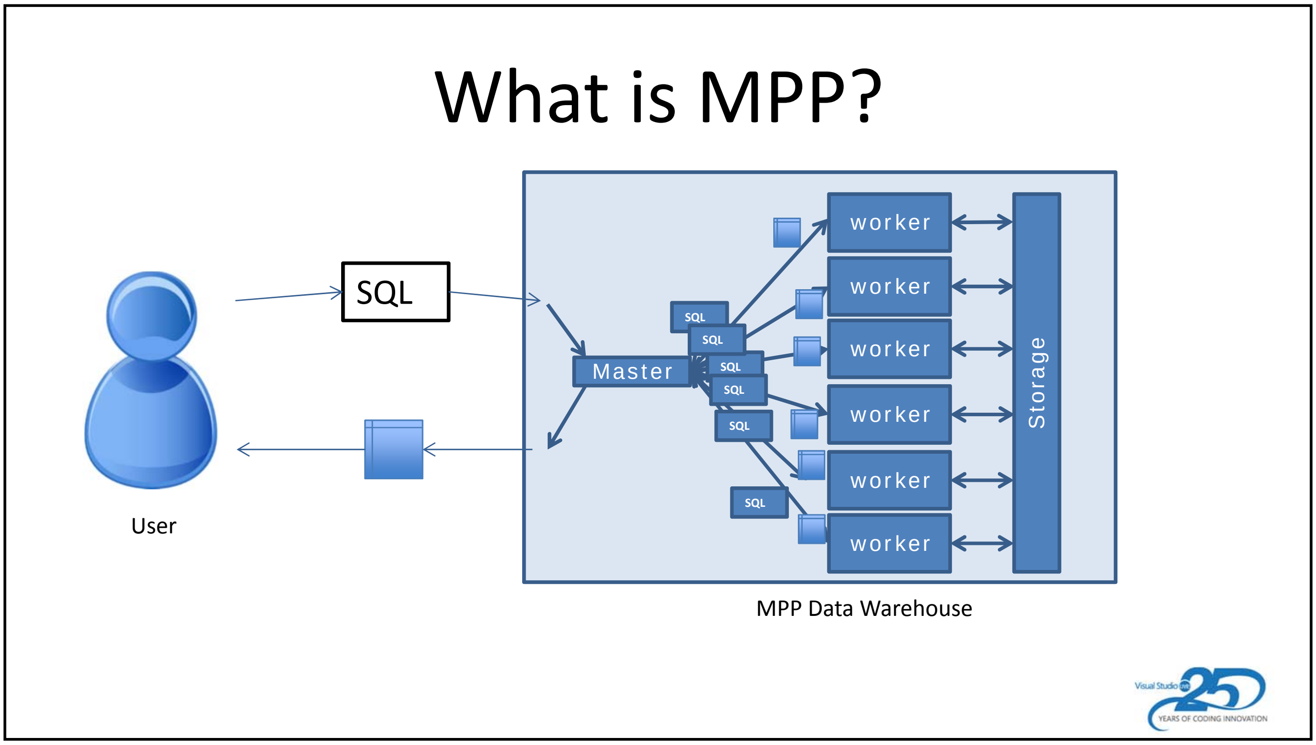
Data Warehouse Query



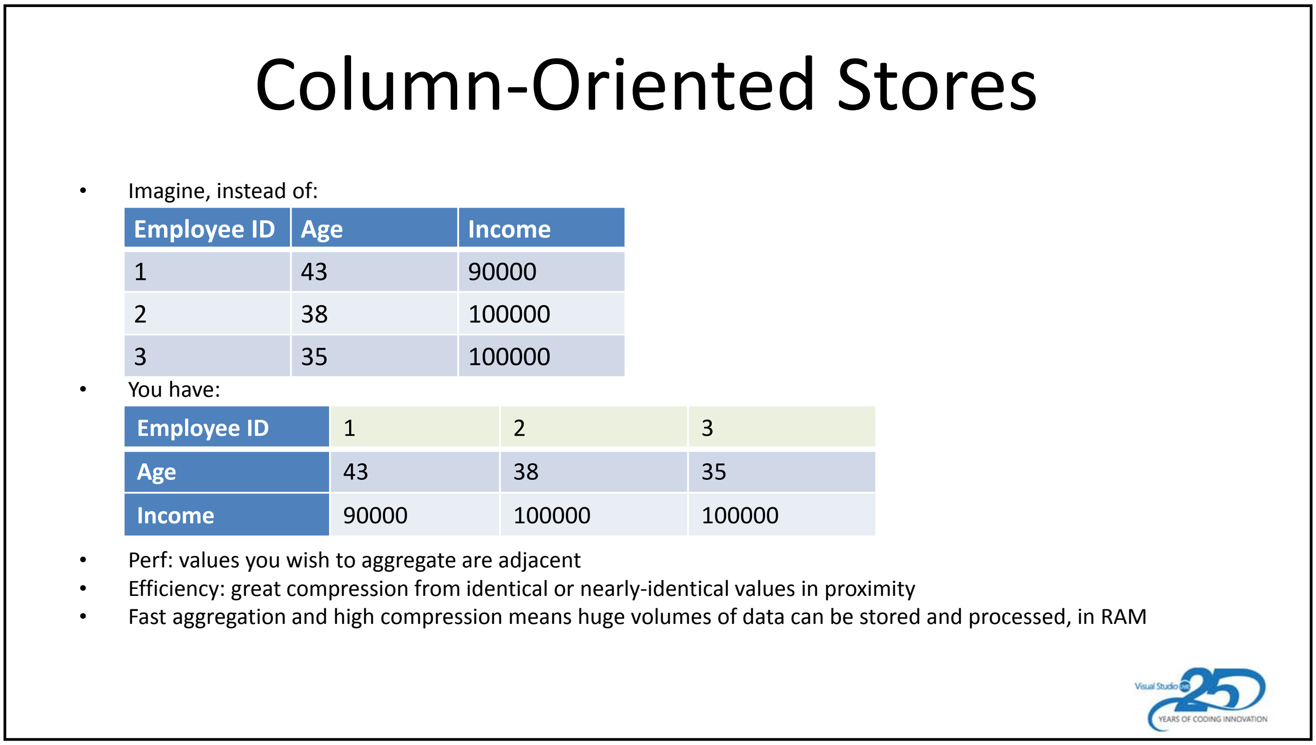
	STATE	Month_Name	(No column name)
1	NM	August 1996	3343.60
2	WY	August 1996	48.00
3	ID	December 1996	6038.60
4	OR	December 1996	780.00
5	WY	December 1996	3391.20
6	NM	July 1996	624.80
7	WA	July 1996	676.00
8	NM	November 1996	1731.20
9	WA	November 1996	2856.00
10	WY	November 1996	141.60
11	AK	October 1996	934.50

What is MPP?

- Massively Parallel Processing
  - A cluster of individual RDBMS instances (worker nodes)
  - One master node, in front
    - Takes query, delegates parts of it to different worker nodes
    - Combines worker nodes' results, returns as single result set
  - Thus, appears as a single RDBMS
    - Send it one query, get back one result set
    - But query is highly parallelized, so it's fast
    - Perfect for data warehouses
  - Bears some resemblance to MapReduce
  - Examples include Teradata, HP Vertica, IBM Netezza, Pivotal Greenplum



What is MPP?



Column-Oriented Stores

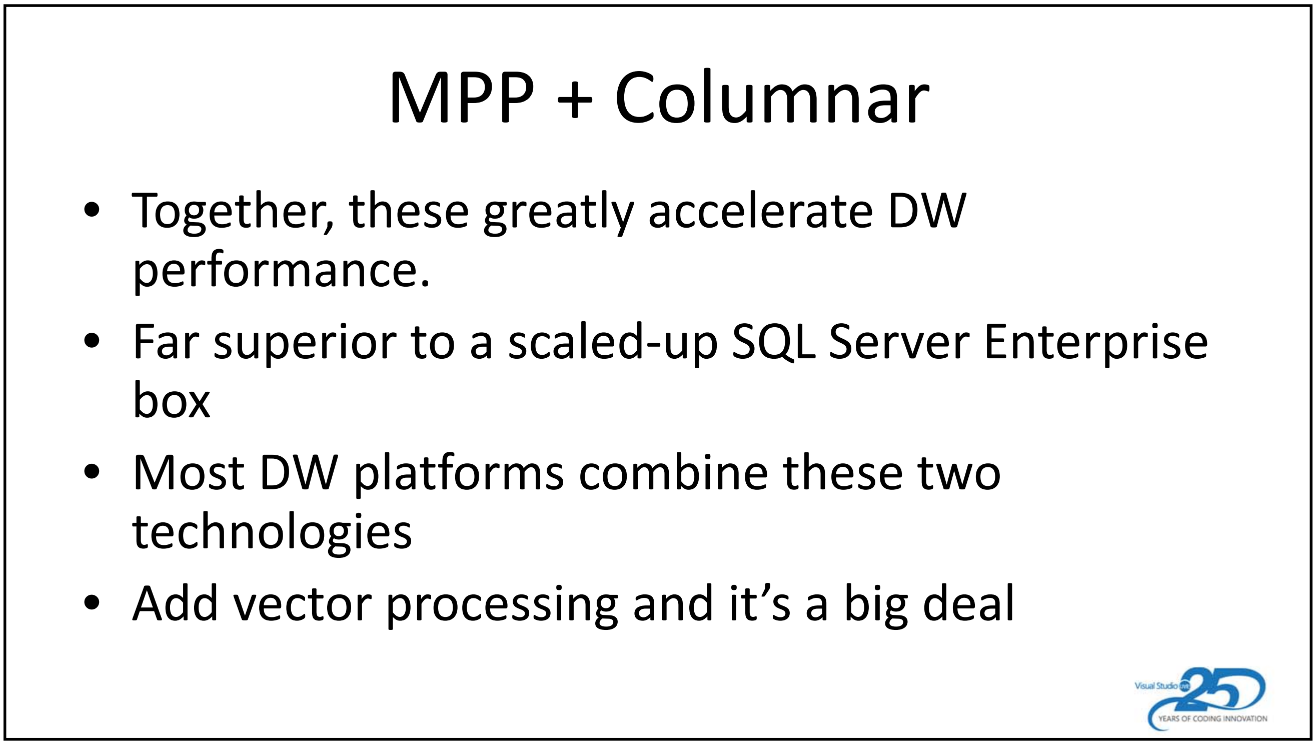
- Imagine, instead of:

Employee ID	Age	Income
1	43	90000
2	38	100000
3	35	100000

- You have:

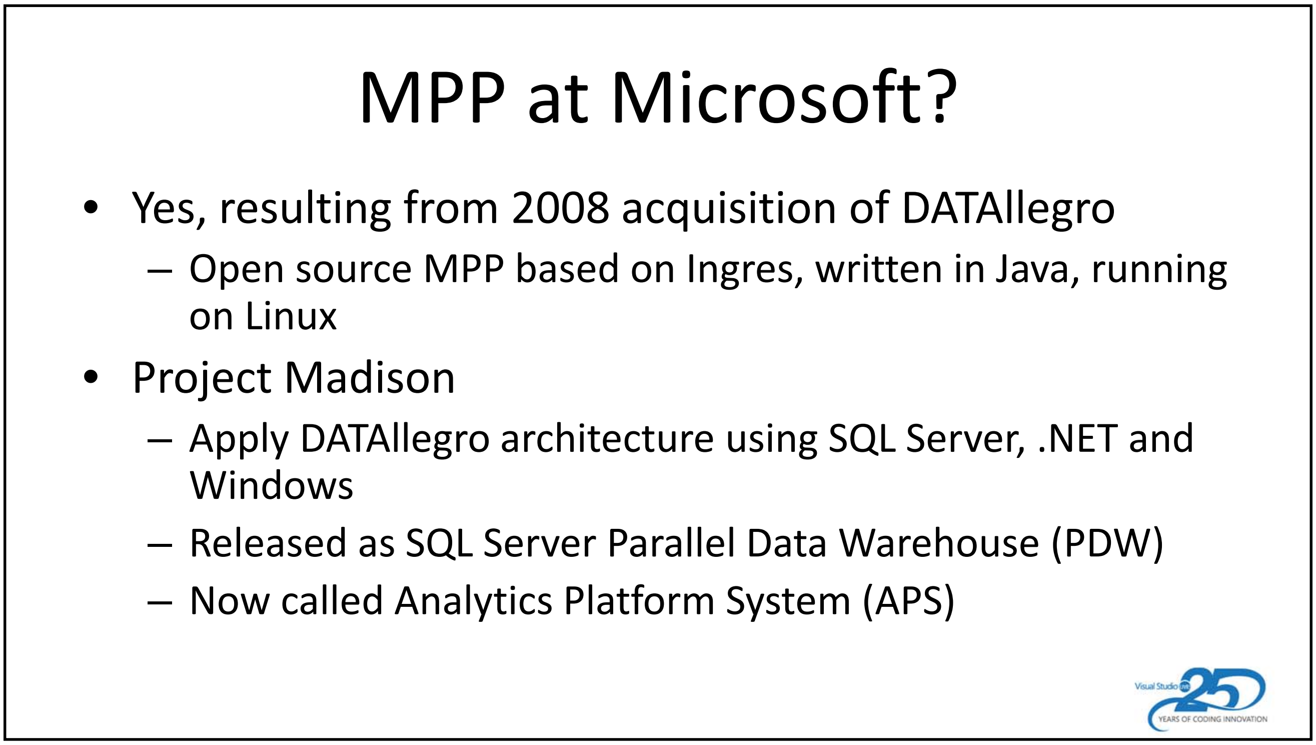
Employee ID	1	2	3
Age	43	38	35
Income	90000	100000	100000

- Perf: values you wish to aggregate are adjacent
- Efficiency: great compression from identical or nearly-identical values in proximity
- Fast aggregation and high compression means huge volumes of data can be stored and processed, in RAM



MPP + Columnar

- Together, these greatly accelerate DW performance.
- Far superior to a scaled-up SQL Server Enterprise box
- Most DW platforms combine these two technologies
- Add vector processing and it’s a big deal



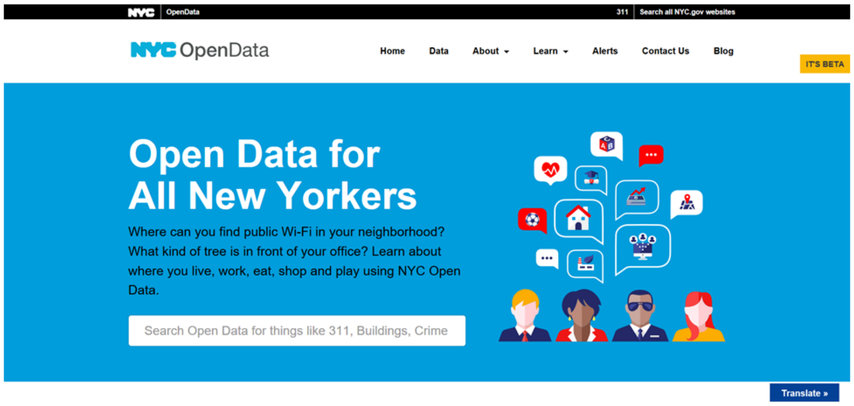
MPP at Microsoft?

- Yes, resulting from 2008 acquisition of DATAllegro
  - Open source MPP based on Ingres, written in Java, running on Linux
- Project Madison
  - Apply DATAllegro architecture using SQL Server, .NET and Windows
  - Released as SQL Server Parallel Data Warehouse (PDW)
  - Now called Analytics Platform System (APS)

NYC 311 Service Calls

OUR DATA SET

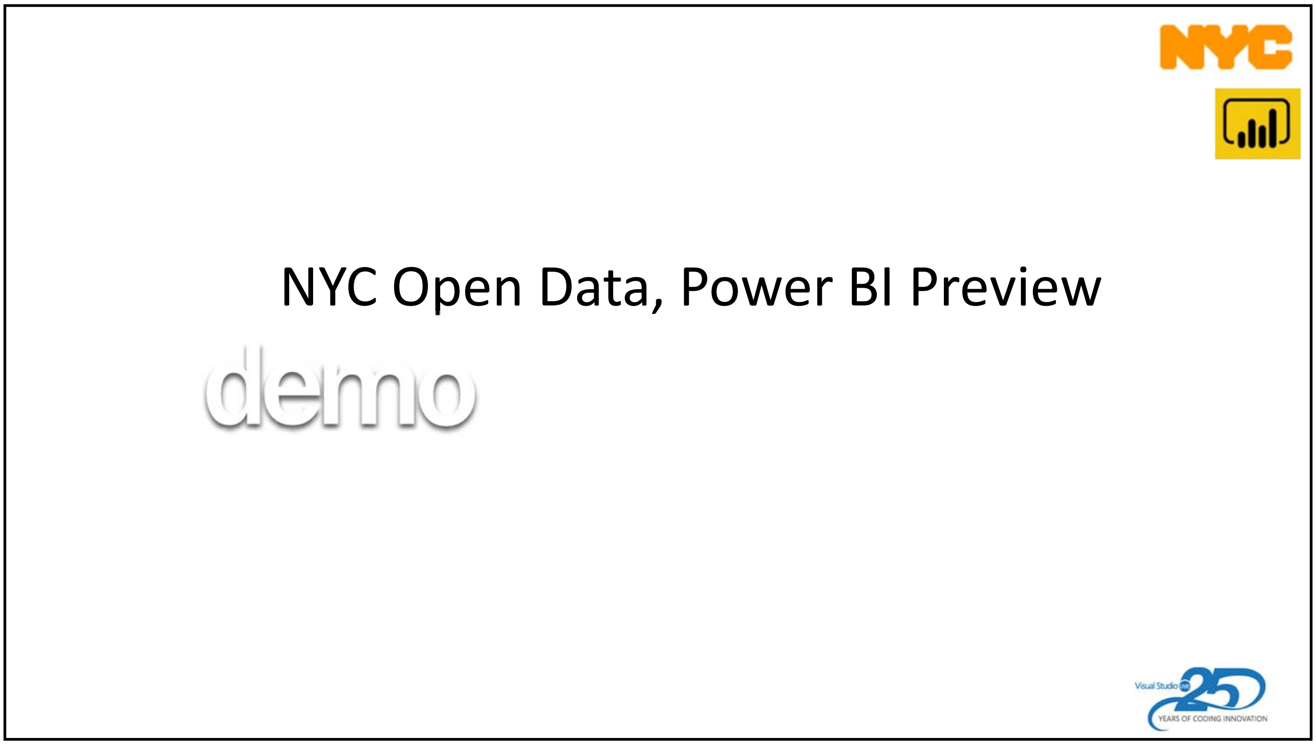
NYC Open Data



Open Data for All New Yorkers

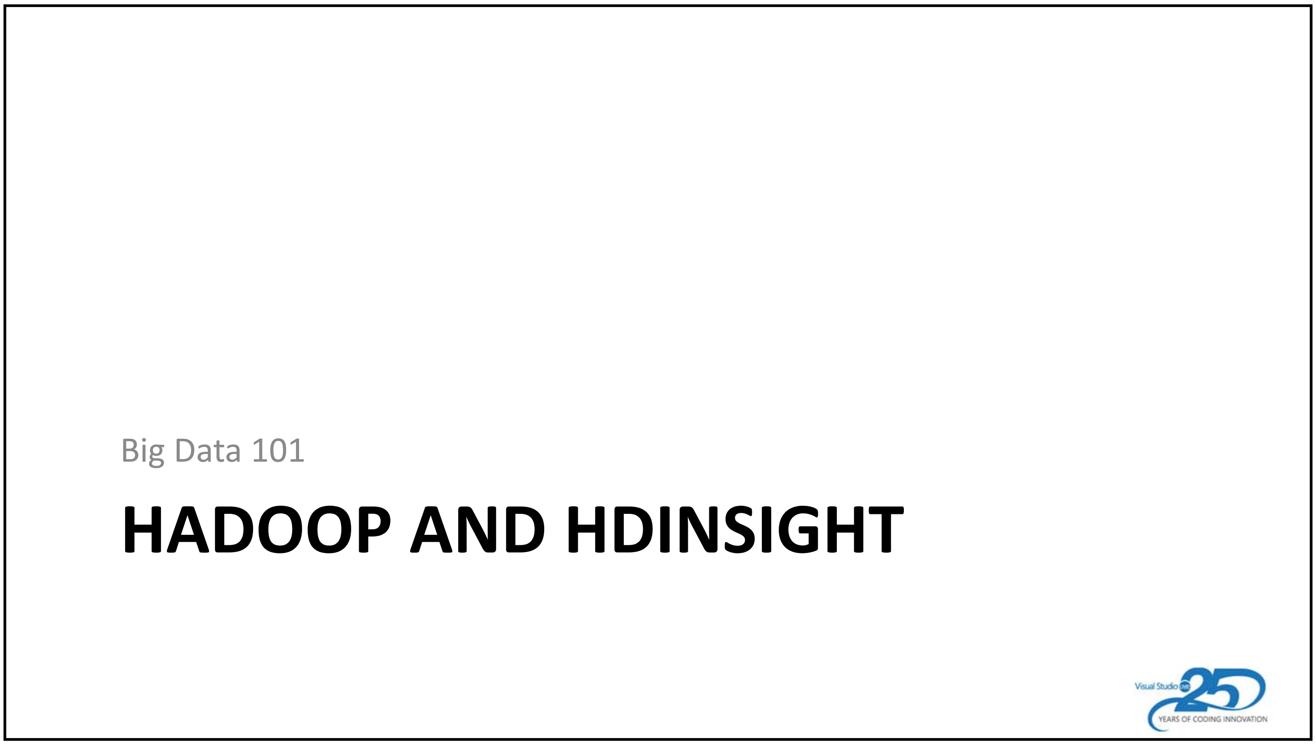
Where can you find public Wi-Fi in your neighborhood? What kind of tree is in front of your office? Learn about where you live, work, eat, shop and play using NYC Open Data.

- https://data.cityofnewyork.us/Social-Services/311-Service-Requests-from-2010-to-Present/erm2-nwe9



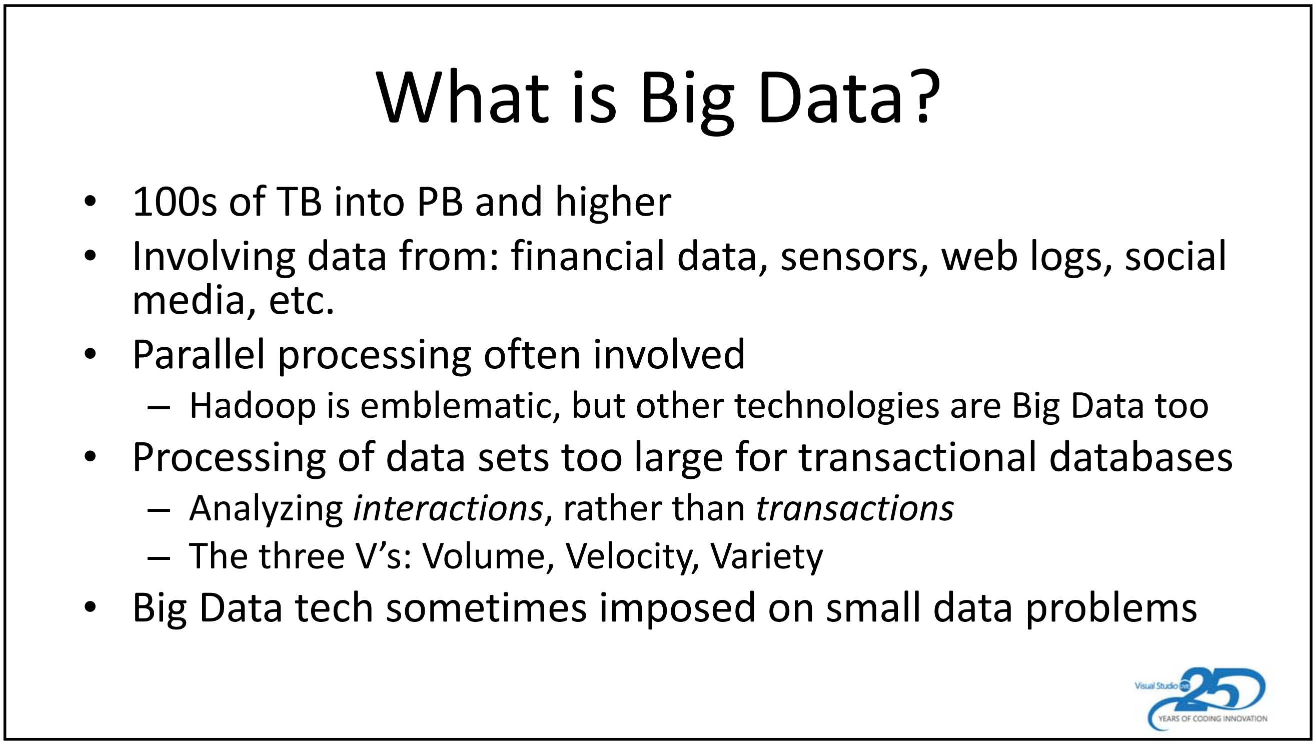
NYC Open Data, Power BI Preview

demo



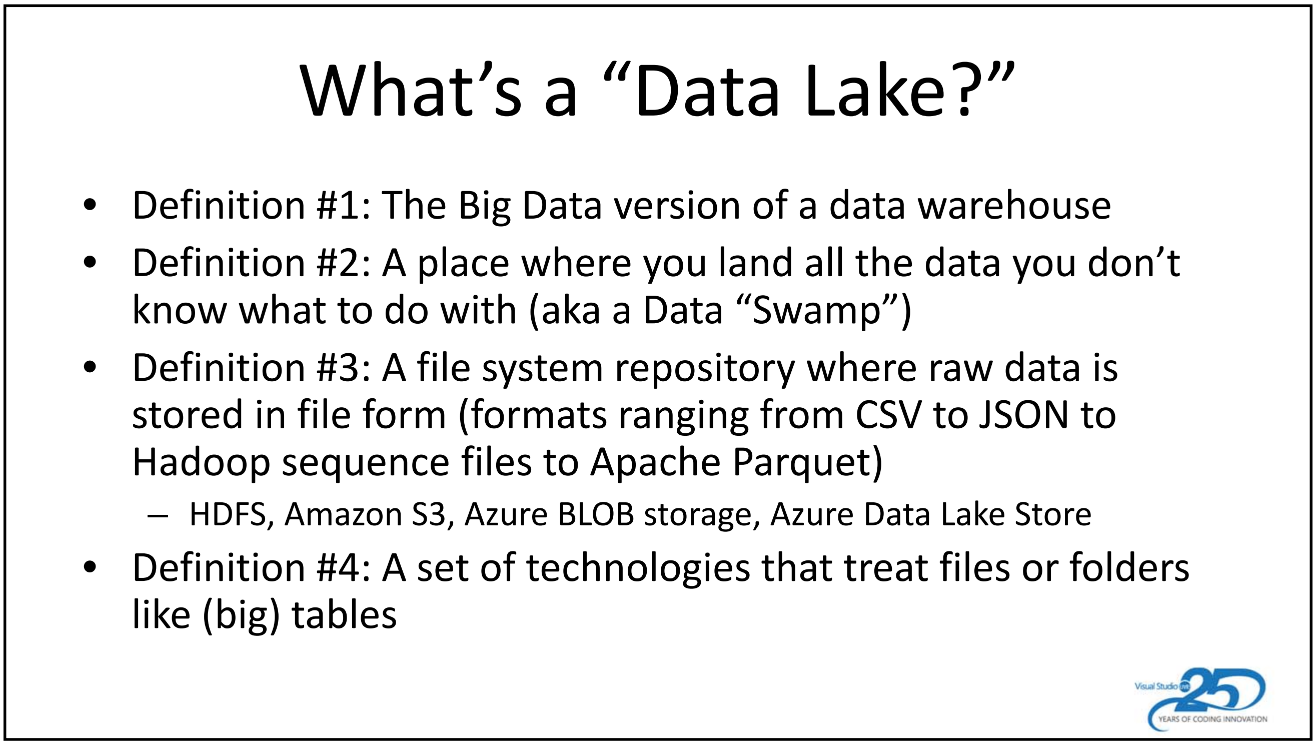
Big Data 101

HADOOP AND HDINSIGHT



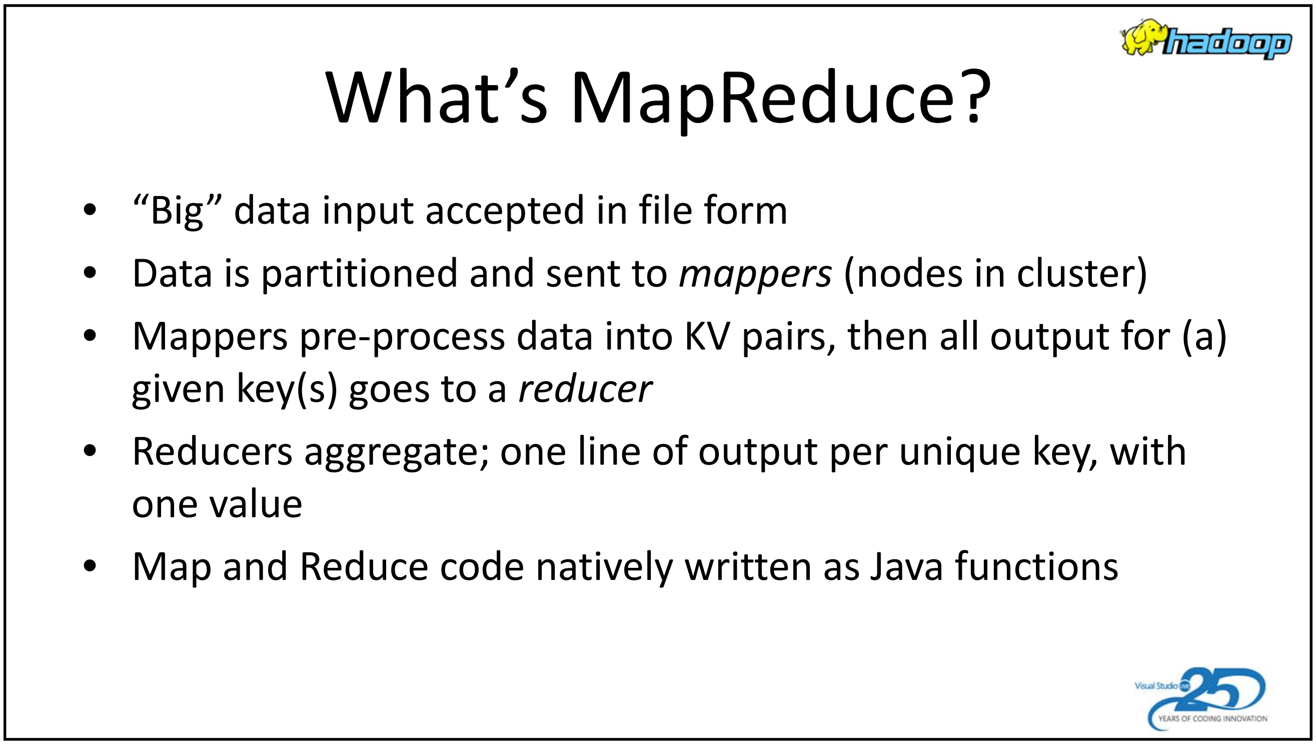
What is Big Data?

- 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
- Big Data tech sometimes imposed on small data problems



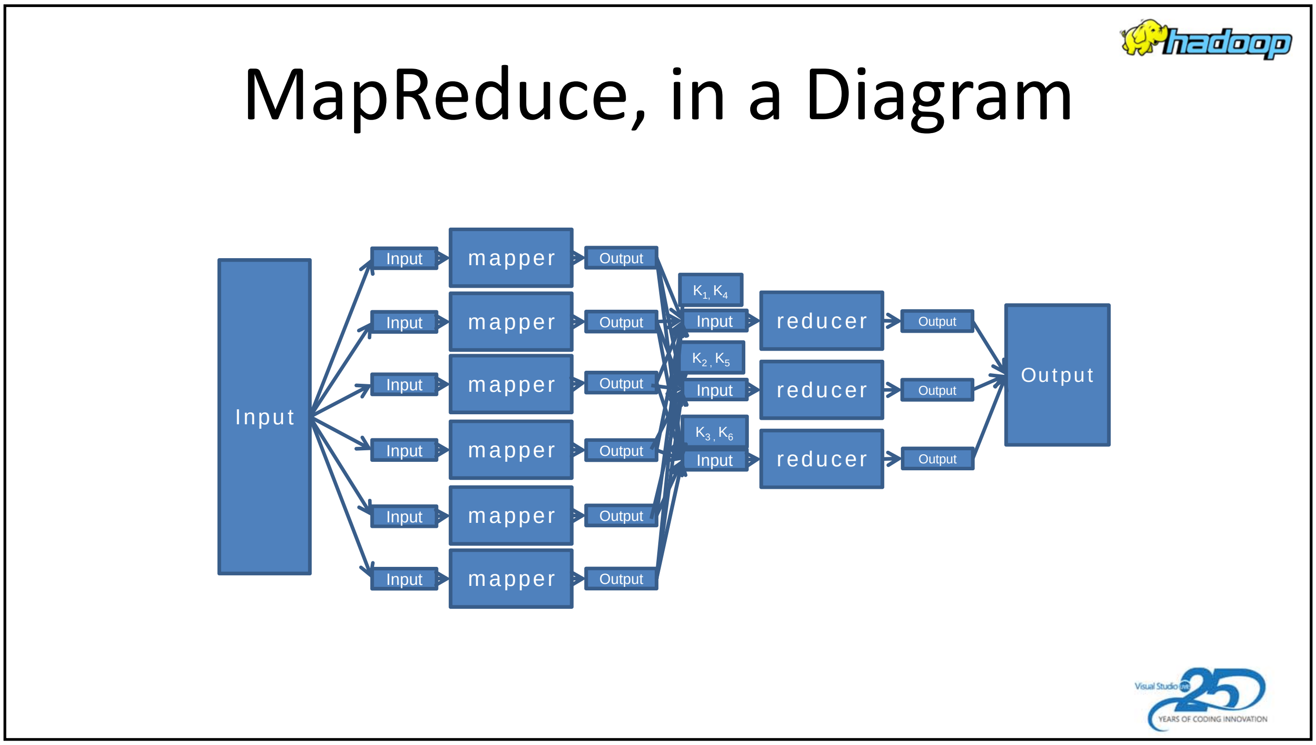
What's a "Data Lake?"

- Definition #1: The Big Data version of a data warehouse
- Definition #2: A place where you land all the data you don't know what to do with (aka a Data "Swamp")
- Definition #3: A file system repository where raw data is stored in file form (formats ranging from CSV to JSON to Hadoop sequence files to Apache Parquet)
  - HDFS, Amazon S3, Azure BLOB storage, Azure Data Lake Store
- Definition #4: A set of technologies that treat files or folders like (big) tables

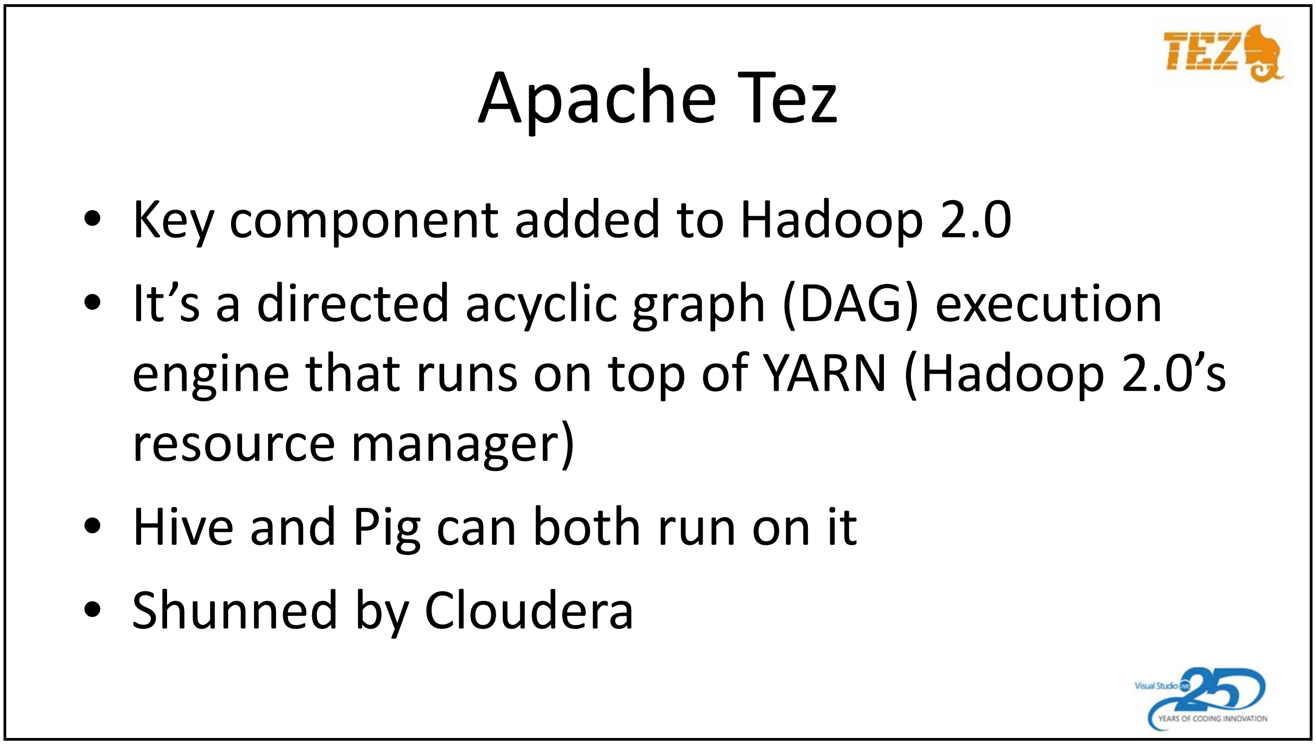


What's MapReduce?

- "Big" data input accepted in file form
- Data is partitioned and sent to *mappers* (nodes in cluster)
- Mappers pre-process data into KV pairs, then all output for (a) given key(s) goes to a *reducer*
- Reducers aggregate; one line of output per unique key, with one value
- Map and Reduce code natively written as Java functions

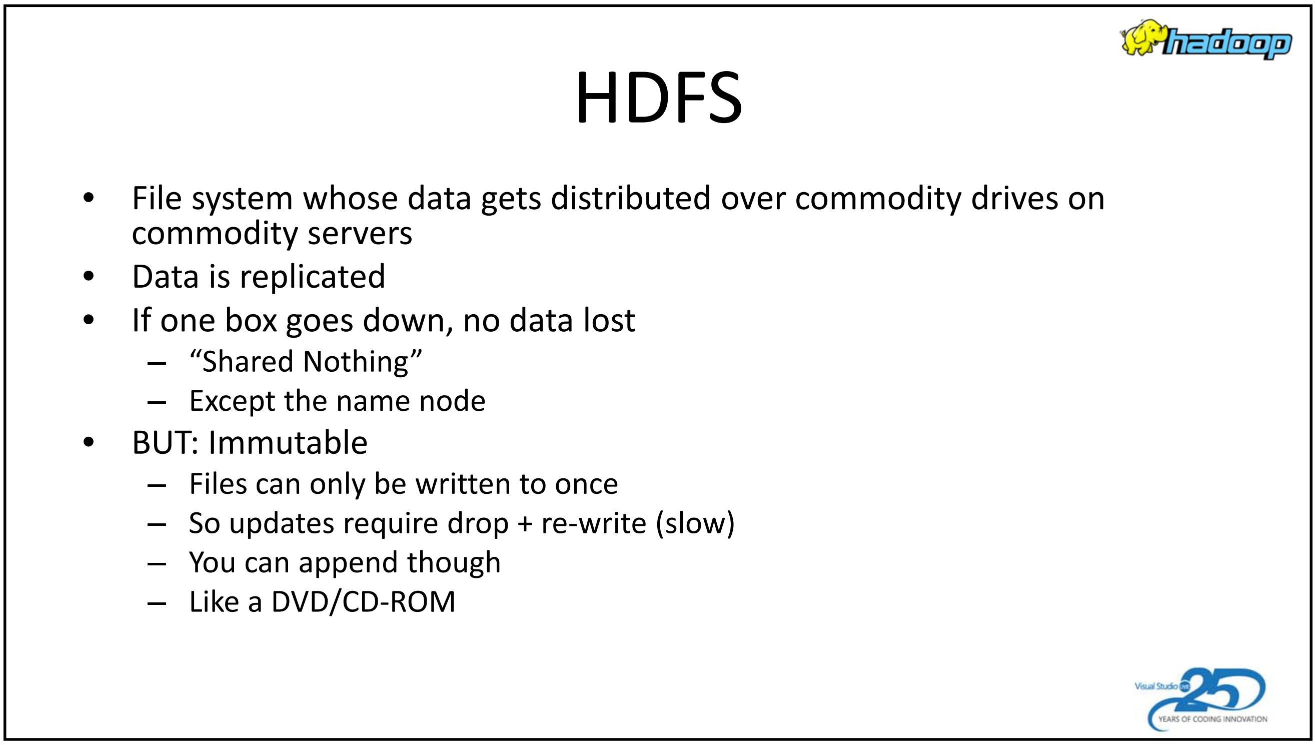


MapReduce, in a Diagram



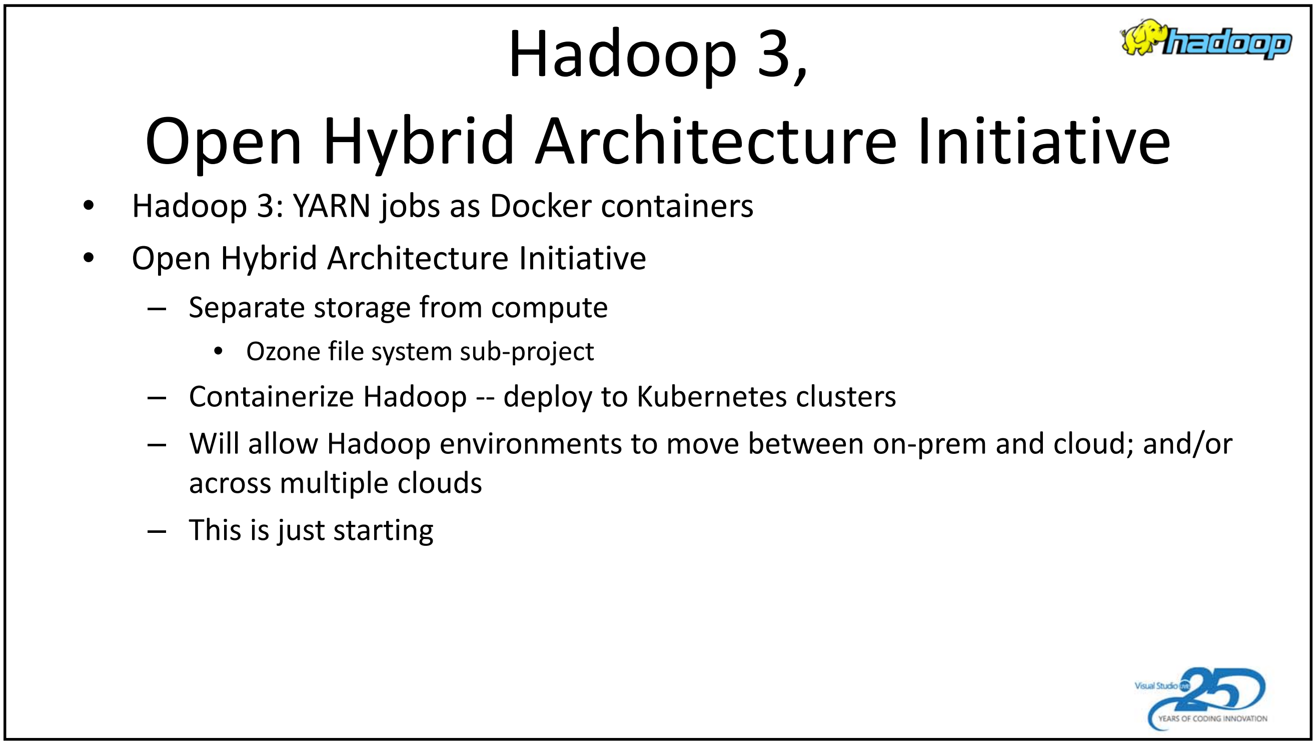
Apache Tez

- Key component added to Hadoop 2.0
- It’s a directed acyclic graph (DAG) execution engine that runs on top of YARN (Hadoop 2.0’s resource manager)
- Hive and Pig can both run on it
- Shunned by Cloudera



HDFS

- File system whose data gets distributed over commodity drives on commodity servers
- Data is replicated
- If one box goes down, no data lost
  - “Shared Nothing”
  - Except the name node
- BUT: Immutable
  - Files can only be written to once
  - So updates require drop + re-write (slow)
  - You can append though
  - Like a DVD/CD-ROM



Hadoop 3, Open Hybrid Architecture Initiative

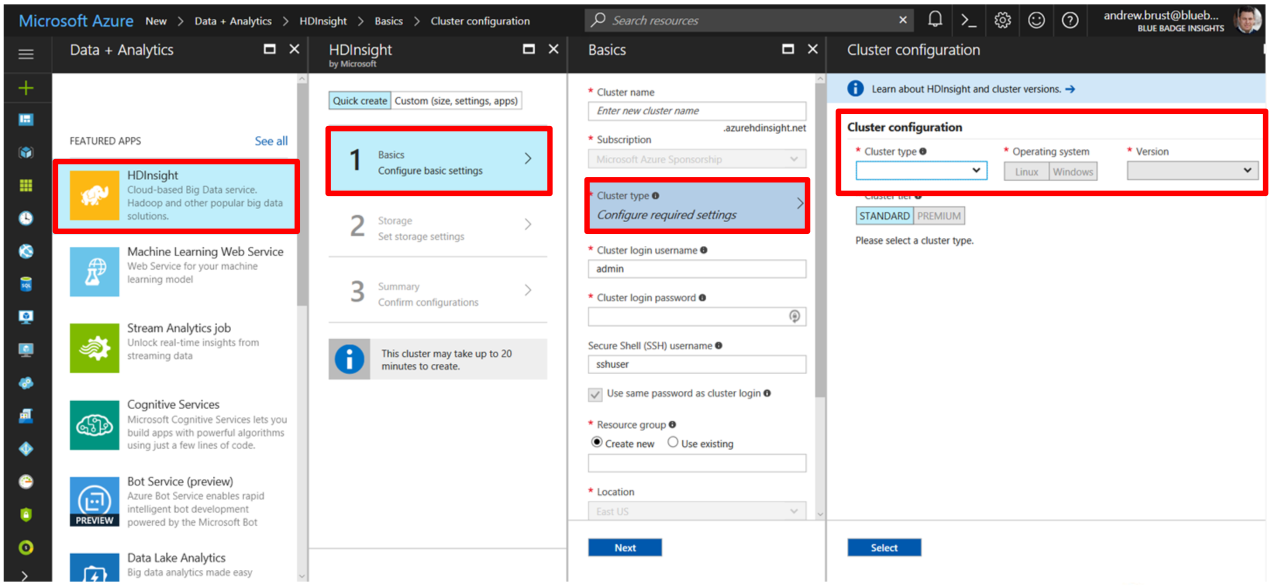
- Hadoop 3: YARN jobs as Docker containers
- Open Hybrid Architecture Initiative
  - Separate storage from compute
    - Ozone file system sub-project
  - Containerize Hadoop -- deploy to Kubernetes clusters
  - Will allow Hadoop environments to move between on-prem and cloud; and/or across multiple clouds
  - This is just starting

HDINSIGHT

Microsoft HDInsight

- Developed with Hortonworks and incorporates Hortonworks Data Platform (HDP) *for Windows*
- Windows Azure HDInsight and Microsoft HDInsight Server
  - Single node preview runs on Windows client
  - Also Hortonworks HDP for Windows
  - Also HDInsight with Analytics Platform System
- Includes ODBC Drivers for Hive
- All contributed back to open source Apache project

Azure HDInsight Provisioning



HDInsight
Cloud-based Big Data service. Hadoop and other popular big data solutions.

Machine Learning Web Service
Web Service for your machine learning model

Stream Analytics job
Unlock real-time insights from streaming data

Cognitive Services
Microsoft Cognitive Services lets you build apps with powerful algorithms using just a few lines of code.

Bot Service (preview)
Azure Bot Service enables rapid intelligent bot development powered by the Microsoft Bot

Data Lake Analytics
Big data analytics made easy

HDInsight
by Microsoft

Quick create | Custom (size, settings, apps)

1 Basics
Configure basic settings

2 Storage
Set storage settings

3 Summary
Confirm configurations

This cluster may take up to 20 minutes to create.

Basics

* Cluster name: Enter new cluster name .azurehdinsight.net

* Subscription: Microsoft Azure Sponsorship

* Cluster type: Configure required settings

* Cluster login username: admin

* Cluster login password:

Secure Shell (SSH) username: sshuser

☑ Use same password as cluster login

* Resource group: ● Create new ○ Use existing

* Location: East US

Next

Cluster configuration

Learn about HDInsight and cluster versions. →

Cluster configuration

* Cluster type:

* Operating system: Linux | Windows

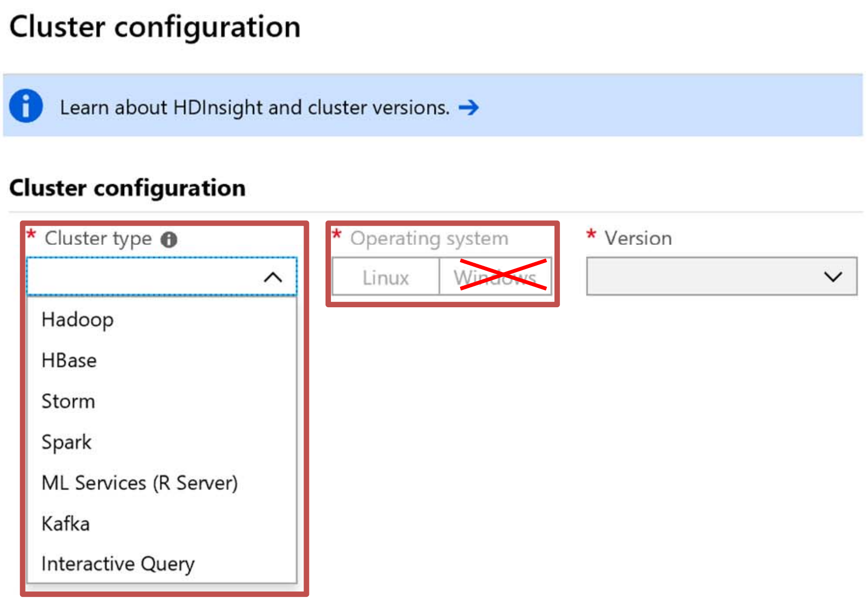
* Version:

Cluster tier: STANDARD | PREMIUM

Please select a cluster type.

Select

Azure HDInsight Provisioning



Cluster configuration

Learn about HDInsight and cluster versions. →

Cluster configuration

* Cluster type

Hadoop
HBase
Storm
Spark
ML Services (R Server)
Kafka
Interactive Query

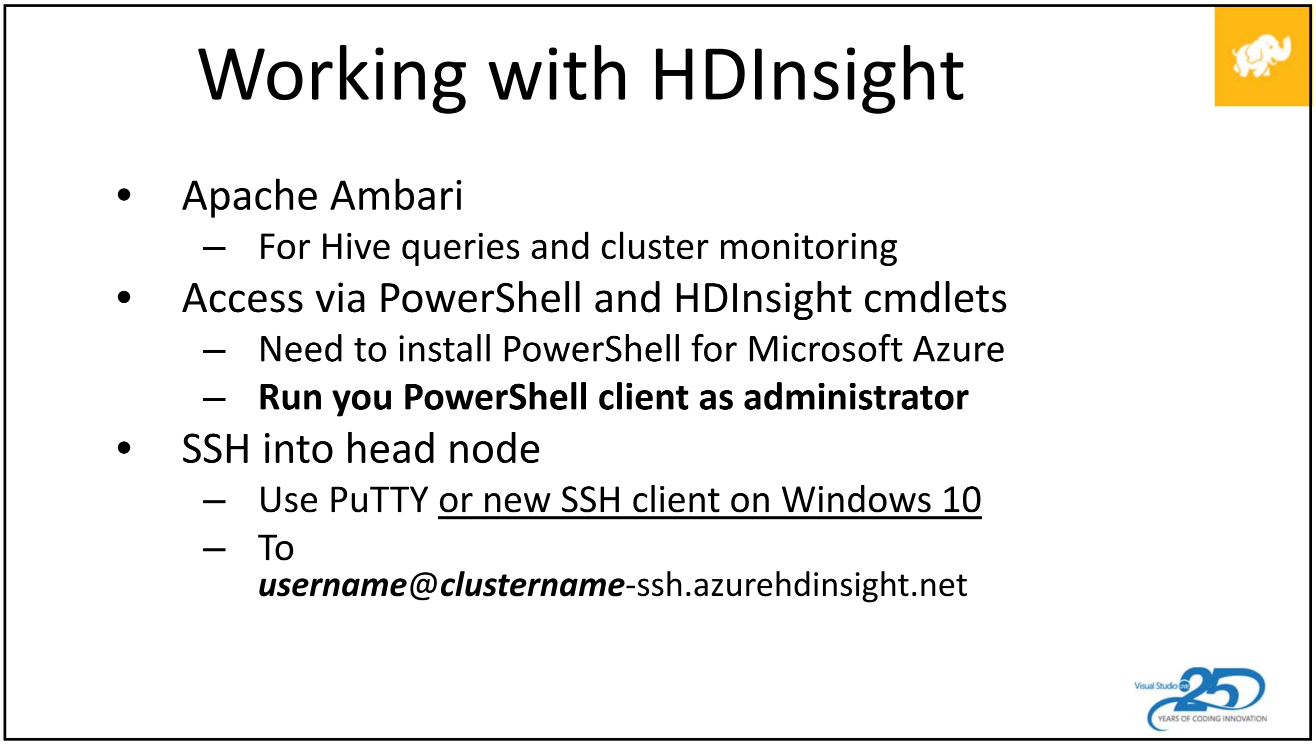
* Operating system

Linux | ~~Windows~~

* Version

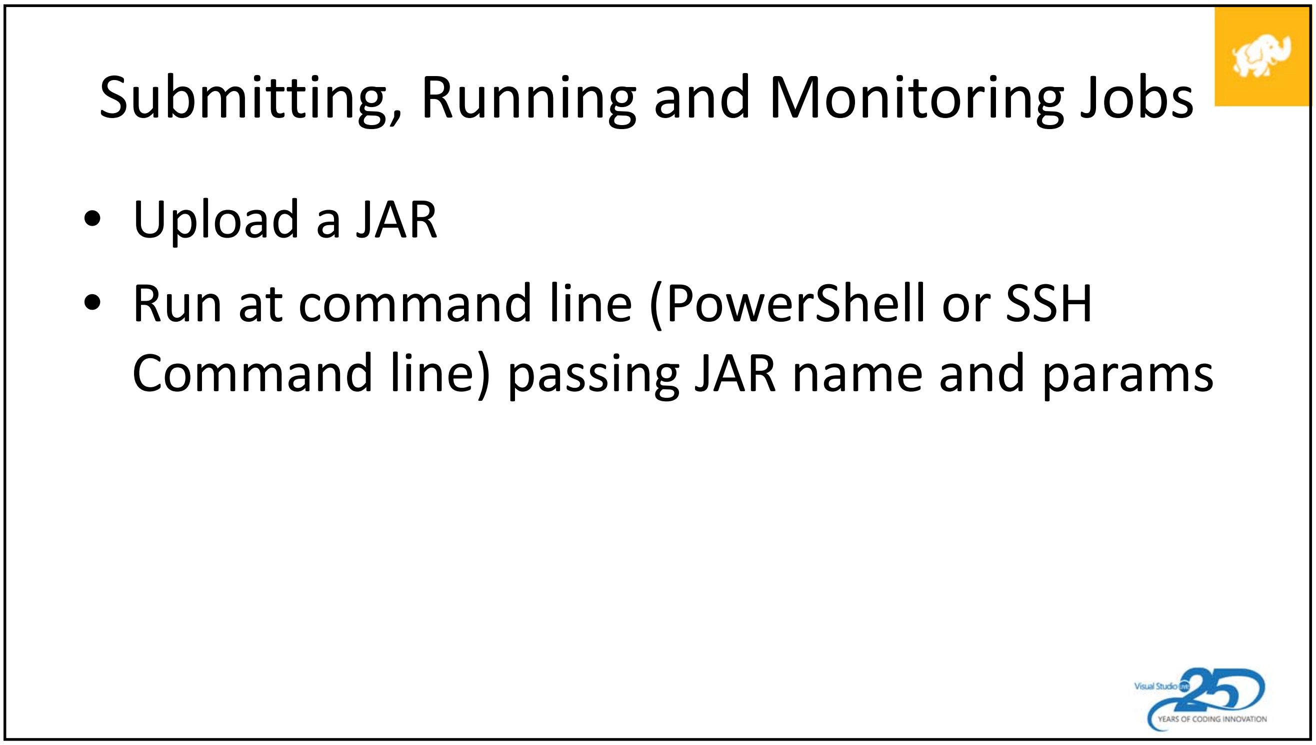
HDInsight Provisioning

demo



Working with HDInsight

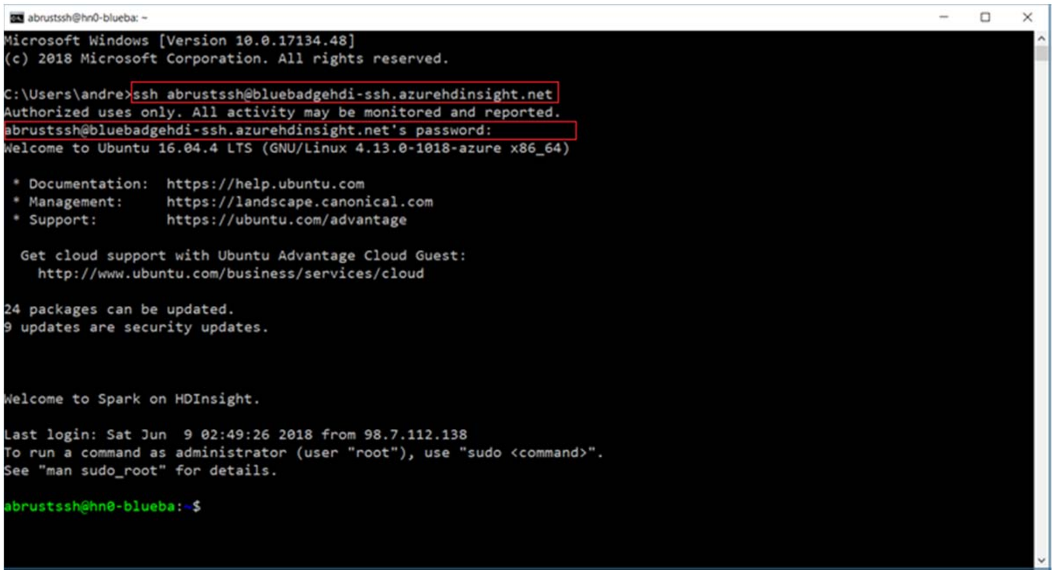
- Apache Ambari
  - For Hive queries and cluster monitoring
- Access via PowerShell and HDInsight cmdlets
  - Need to install PowerShell for Microsoft Azure
  - **Run you PowerShell client as administrator**
- SSH into head node
  - Use PuTTY or new SSH client on Windows 10
  - To ***username@clustername***-ssh.azurehdinsight.net



Submitting, Running and Monitoring Jobs

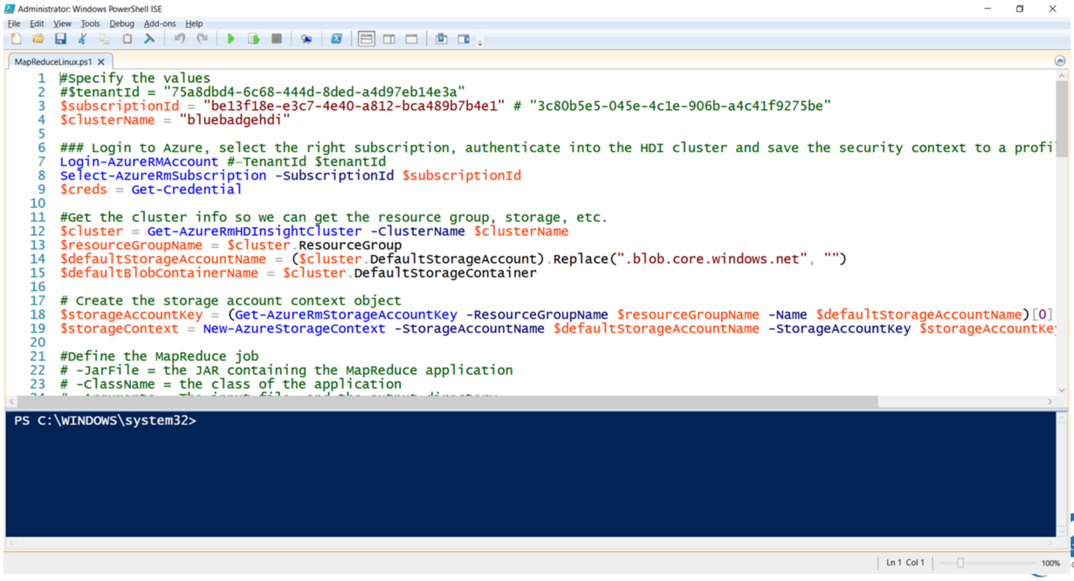
- Upload a JAR
- Run at command line (PowerShell or SSH Command line) passing JAR name and params

Clients Options: Command Line via SSH



```
Microsoft Windows [Version 10.0.17134.48]
(c) 2018 Microsoft Corporation. All rights reserved.

C:\Users\andre>ssh abrustssh@bluebadgehdi-ssh.azurehdinsight.net
Authorized uses only. All activity may be monitored and reported.
abrustssh@bluebadgehdi-ssh.azurehdinsight.net's password:
Welcome to Ubuntu 16.04.4 LTS (GNU/Linux 4.13.0-1018-azure x86_64)

 * Documentation:  https://help.ubuntu.com
 * Management:     https://landscape.canonical.com
 * Support:        https://ubuntu.com/advantage

  Get cloud support with Ubuntu Advantage Cloud Guest:
    http://www.ubuntu.com/business/services/cloud

24 packages can be updated.
9 updates are security updates.


Welcome to Spark on HDInsight.

Last login: Sat Jun  9 02:49:26 2018 from 98.7.112.138
To run a command as administrator (user "root"), use "sudo <command>".
See "man sudo_root" for details.

abrustssh@hn0-blueba:~$
```

Clients Options: PowerShell

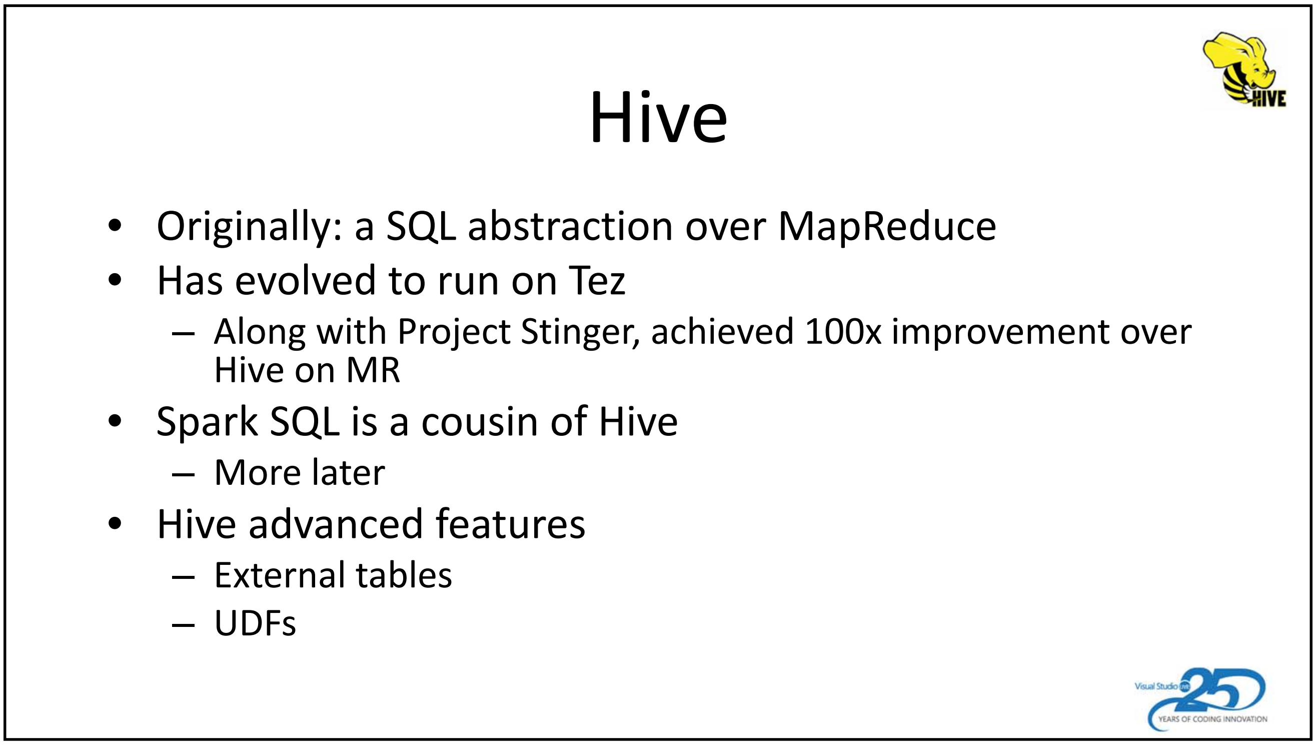

Administrator: Windows PowerShell ISE
File Edit View Tools Debug Add-ons Help
MapReduceLinux.ps1
Ln 1 Col 1
100%

```
#Specify the values
#$tenantId = "75a8dbd4-6c68-444d-8ded-a4d97eb14e3a"
$subscriptionId = "be13f18e-e3c7-4e40-a812-bca489b7b4e1" # "3c80b5e5-045e-4c1e-906b-a4c41f9275be"
$clusterName = "bluebadgehdi"

### Login to Azure, select the right subscription, authenticate into the HDI cluster and save the security context to a profi
Login-AzureRMAccount #-TenantId $tenantId
Select-AzureRmSubscription -SubscriptionId $subscriptionId
$creds = Get-Credential

#Get the cluster info so we can get the resource group, storage, etc.
$cluster = Get-AzureRmHDInsightCluster -ClusterName $clusterName
$resourceGroupName = $cluster.ResourceGroup
$defaultStorageAccountName = ($cluster.DefaultStorageAccount).Replace(".blob.core.windows.net", "")
$defaultBlobContainerName = $cluster.DefaultStorageContainer

# Create the storage account context object
$storageAccountKey = (Get-AzureRmStorageAccountKey -ResourceGroupName $resourceGroupName -Name $defaultStorageAccountName)[0]
$storageContext = New-AzureStorageContext -StorageAccountName $defaultStorageAccountName -StorageAccountKey $storageAccountKe

#Define the MapReduce job
# -JarFile = the JAR containing the MapReduce application
# -ClassName = the class of the application
```

```
PS C:\WINDOWS\system32>
```

WordCount Code;
Running MapReduce Jobs

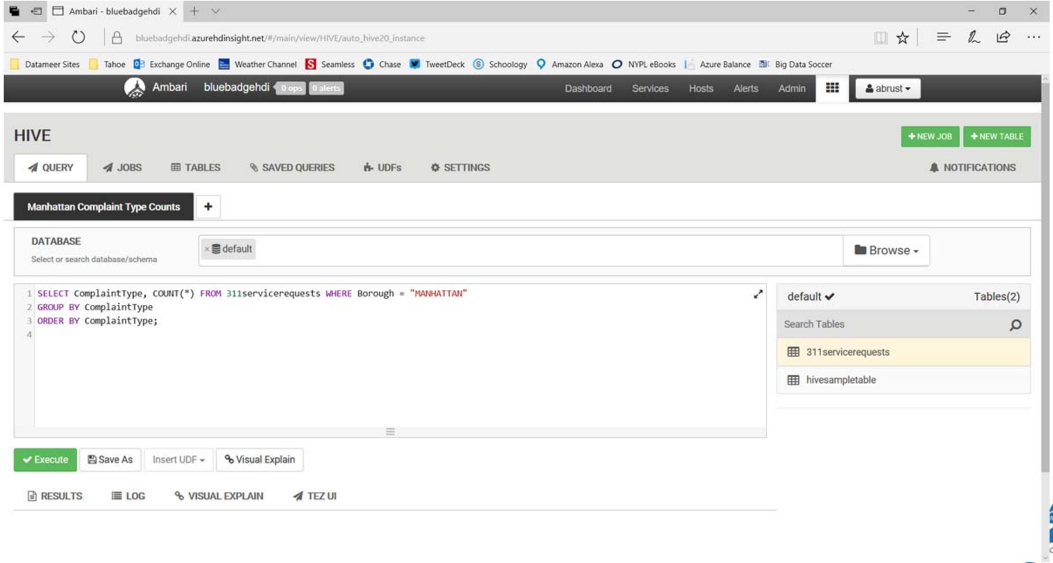
demo



Hive

- Originally: a SQL abstraction over MapReduce
- Has evolved to run on Tez
  - Along with Project Stinger, achieved 100x improvement over Hive on MR
- Spark SQL is a cousin of Hive
  - More later
- Hive advanced features
  - External tables
  - UDFs

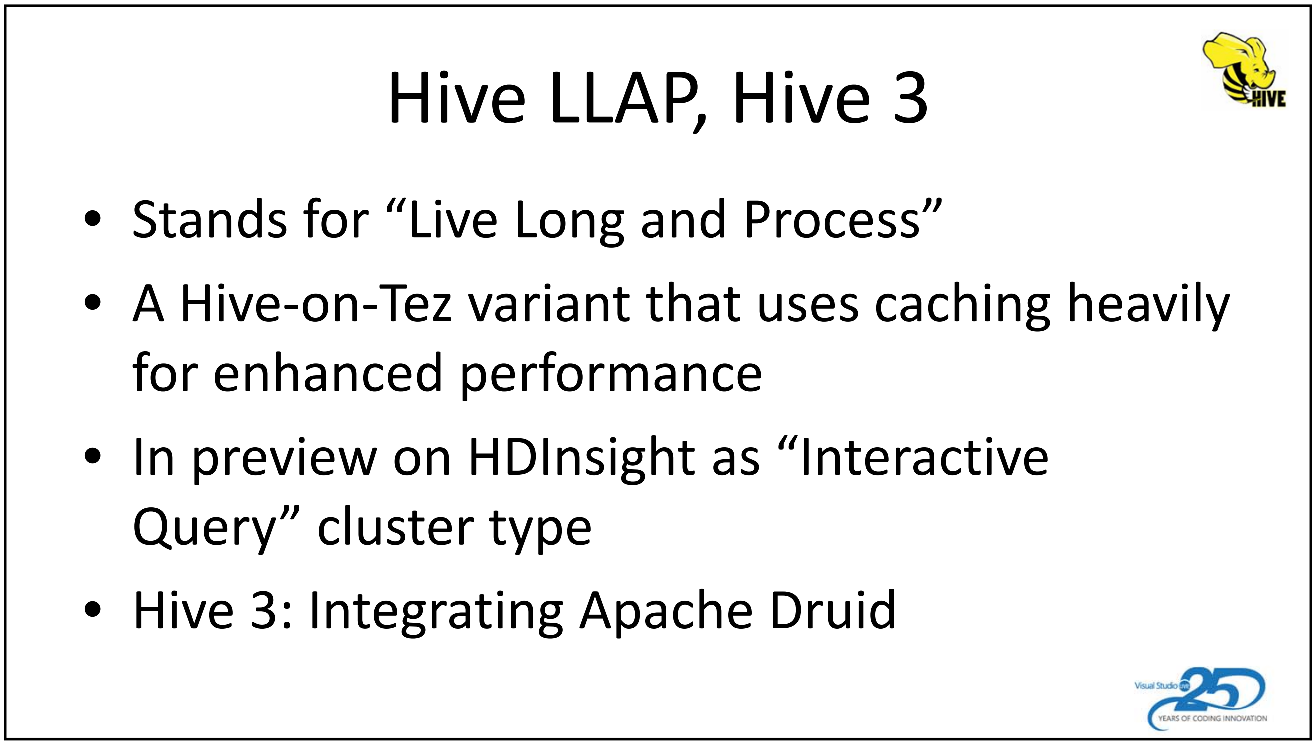
Clients Options: Ambari



```sql
SELECT ComplaintType, COUNT(*) FROM 311servicerequests WHERE Borough = "MANHATTAN"
GROUP BY ComplaintType
ORDER BY ComplaintType;
```

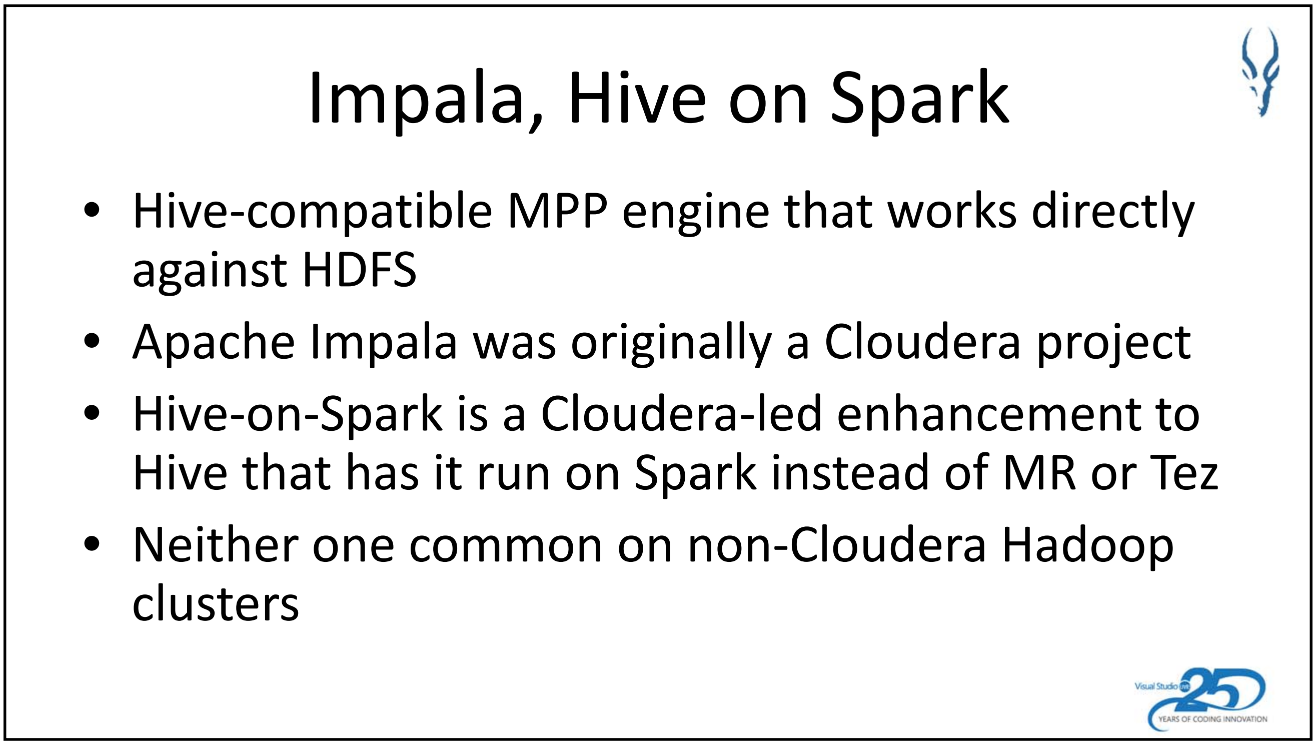
Hive

demo



Hive LLAP, Hive 3

- Stands for “Live Long and Process”
- A Hive-on-Tez variant that uses caching heavily for enhanced performance
- In preview on HDInsight as “Interactive Query” cluster type
- Hive 3: Integrating Apache Druid



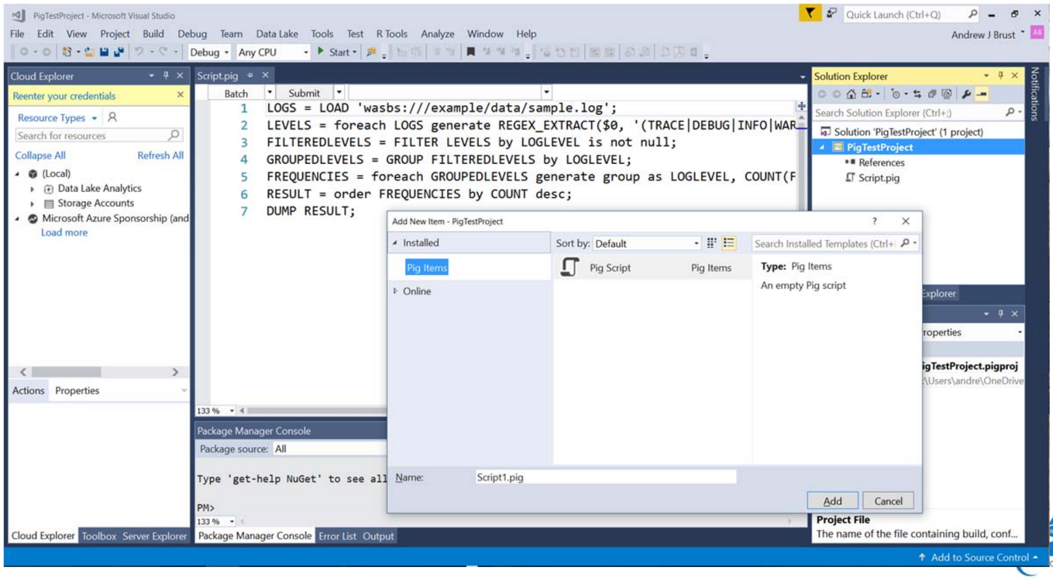
Impala, Hive on Spark

- Hive-compatible MPP engine that works directly against HDFS
- Apache Impala was originally a Cloudera project
- Hive-on-Spark is a Cloudera-led enhancement to Hive that has it run on Spark instead of MR or Tez
- Neither one common on non-Cloudera Hadoop clusters

Pig

- Also a programming language abstraction over MapReduce
- Language is called Pig Latin
- Can be used for interactively and for queries
- More often used for data transformation, from scripts

Clients Options: Azure Data Lake Tools in VS



```
LOGS = LOAD 'wasbs:///example/data/sample.log';
LEVELS = foreach LOGS generate REGEX_EXTRACT($0, '(TRACE|DEBUG|INFO|WAR
FILTEREDLEVELS = FILTER LEVELS by LOGLEVEL is not null;
GROUPEDLEVELS = GROUP FILTEREDLEVELS by LOGLEVEL;
FREQUENCIES = foreach GROUPEDLEVELS generate group as LOGLEVEL, COUNT(F
RESULT = order FREQUENCIES by COUNT desc;
DUMP RESULT;
```

Type: Pig Items

An empty Pig script

Name: Script1.pig

Package source: All

```
Type 'get-help NuGet' to see all

PM>
```

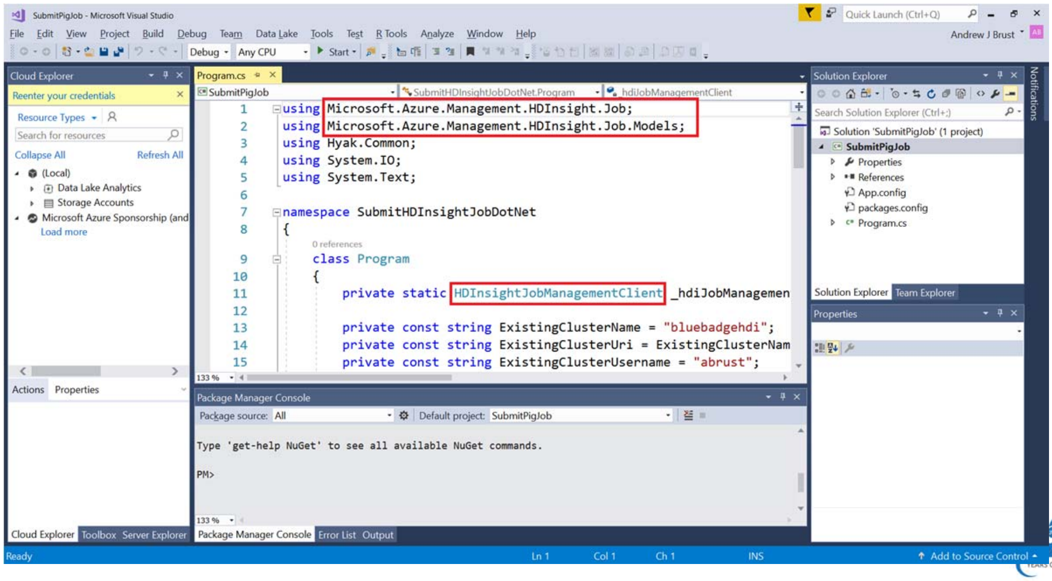
Project File

The name of the file containing build, conf...

igTestProject.pigproj

\Users\andre\OneDrive

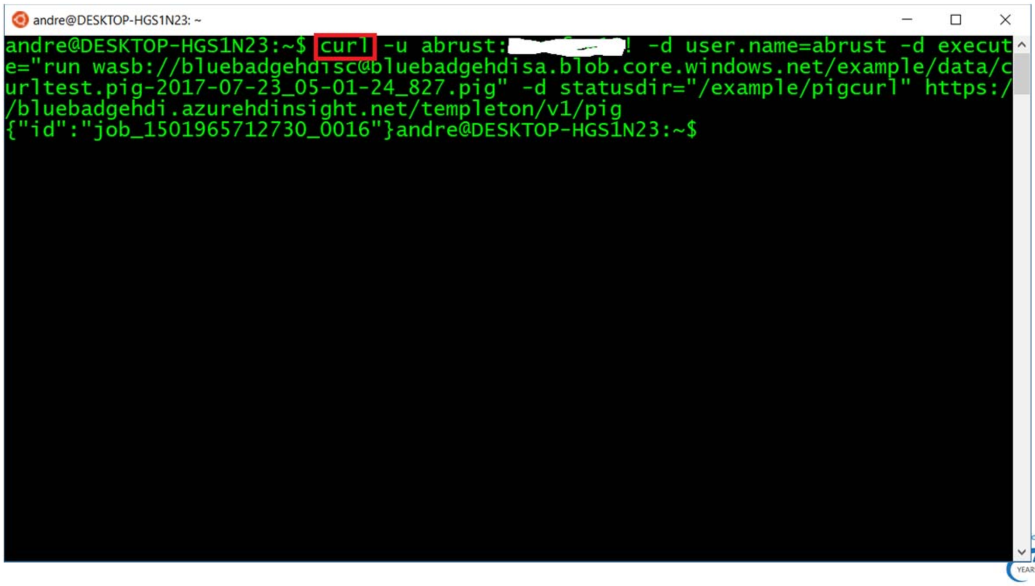
Clients Options: .NET SDK for Hadoop



```
using Microsoft.Azure.Management.HDInsight.Job;
using Microsoft.Azure.Management.HDInsight.Job.Models;
using Hyak.Common;
using System.IO;
using System.Text;

namespace SubmitHDInsightJobDotNet
{
    0 references
    class Program
    {
        private static HDInsightJobManagementClient _hdiJobManagemen

        private const string ExistingClusterName = "bluebadgehdi";
        private const string ExistingClusterUri = ExistingClusterNam
        private const string ExistingClusterUsername = "abrust";
```

```
Type 'get-help NuGet' to see all available NuGet commands.

PM>
```

Clients Options: cURL

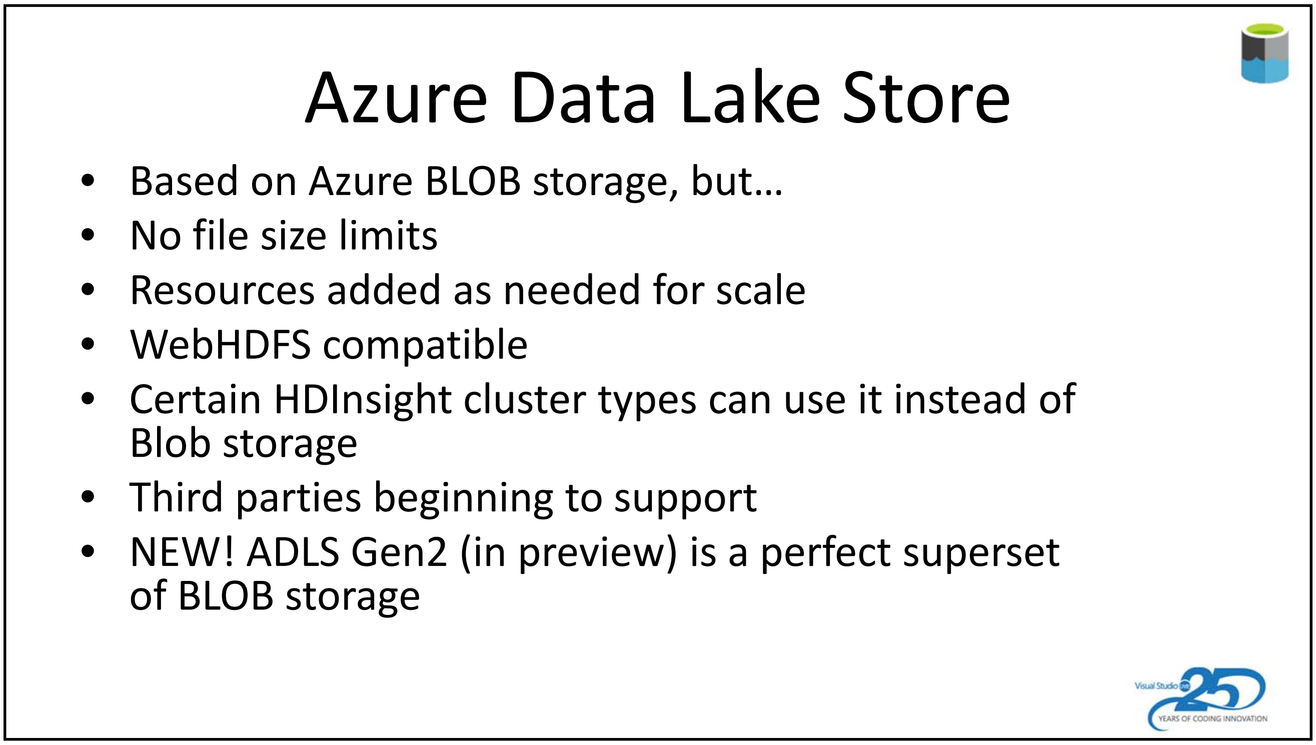


```
andre@DESKTOP-HGS1N23:~$ curl -u abrust:[redacted] -d user.name=abrust -d execut
e="run wasb://bluebadgehdisc@bluebadgehdisa.blob.core.windows.net/example/data/c
urltest.pig-2017-07-23_05-01-24_827.pig" -d statusdir="/example/pigcurl" https:/
/bluebadgehdi.azurehdinsight.net/templeton/v1/pig
{"id":"job_1501965712730_0016"}andre@DESKTOP-HGS1N23:~$
```

Pig

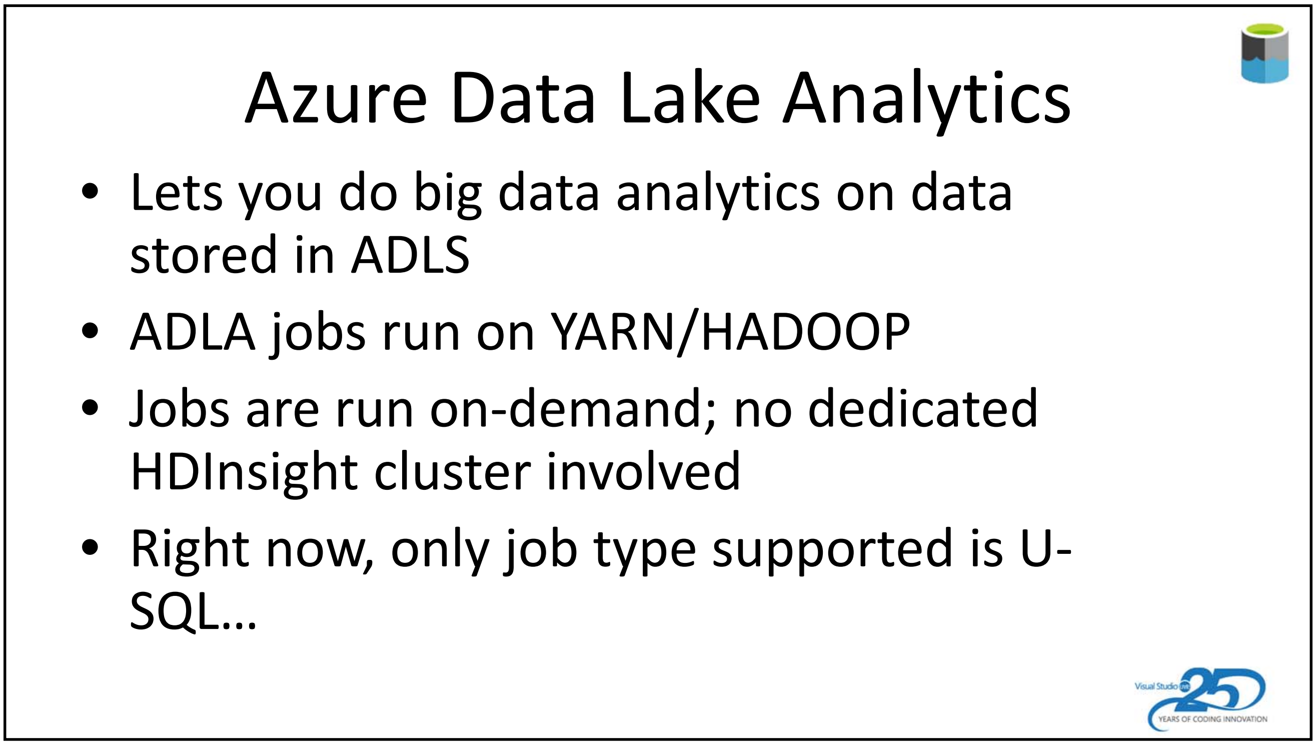
demo

AZURE DATA LAKE



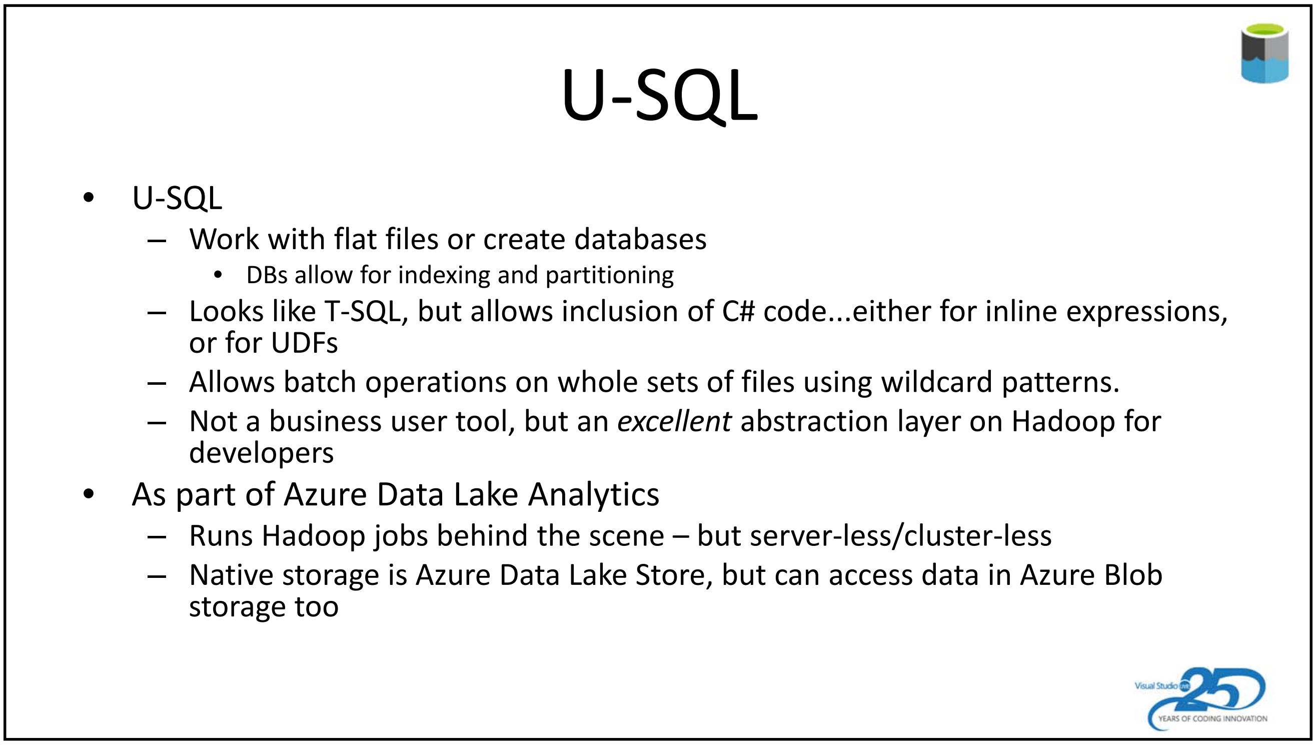
Azure Data Lake Store

- Based on Azure BLOB storage, but…
- No file size limits
- Resources added as needed for scale
- WebHDFS compatible
- Certain HDInsight cluster types can use it instead of Blob storage
- Third parties beginning to support
- NEW! ADLS Gen2 (in preview) is a perfect superset of BLOB storage



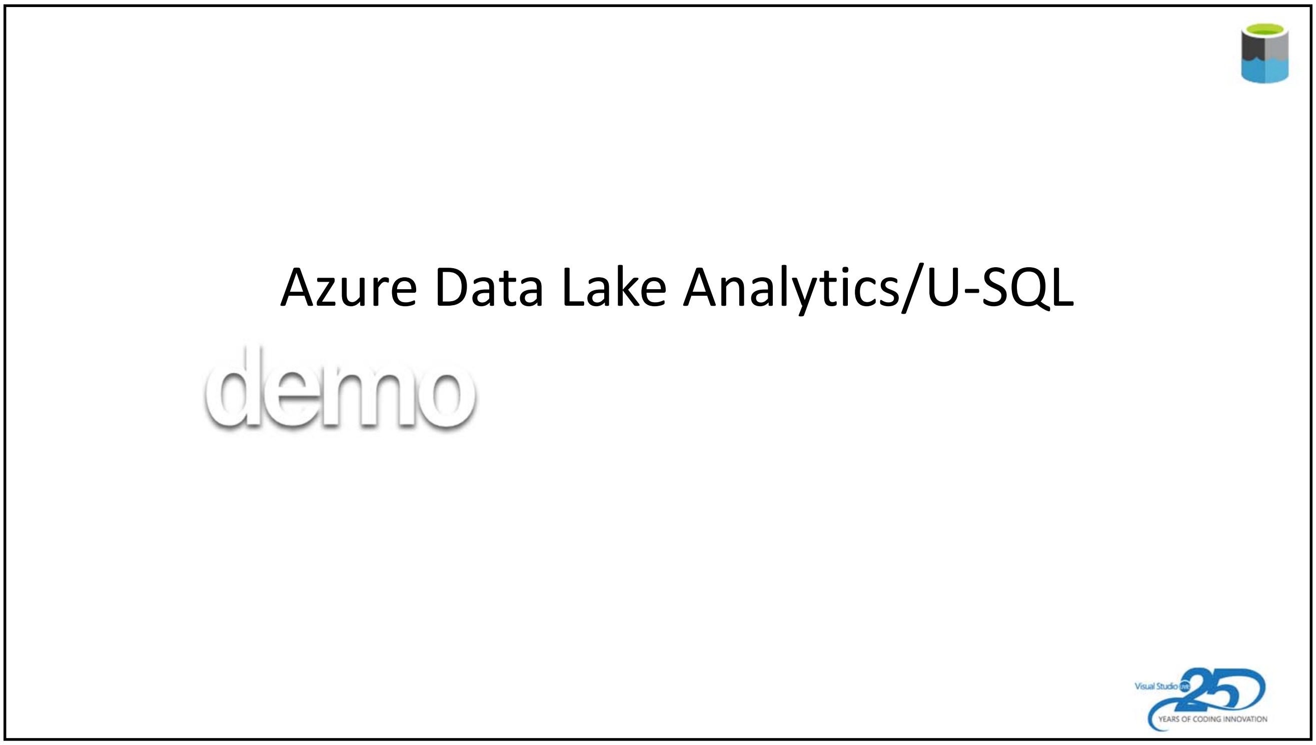
Azure Data Lake Analytics

- Lets you do big data analytics on data stored in ADLS
- ADLA jobs run on YARN/HADOOP
- Jobs are run on-demand; no dedicated HDInsight cluster involved
- Right now, only job type supported is U-SQL…



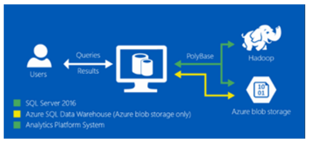
U-SQL

- U-SQL
  - Work with flat files or create databases
    - DBs allow for indexing and partitioning
  - Looks like T-SQL, but allows inclusion of C# code...either for inline expressions, or for UDFs
  - Allows batch operations on whole sets of files using wildcard patterns.
  - Not a business user tool, but an *excellent* abstraction layer on Hadoop for developers
- As part of Azure Data Lake Analytics
  - Runs Hadoop jobs behind the scene – but server-less/cluster-less
  - Native storage is Azure Data Lake Store, but can access data in Azure Blob storage too



Azure Data Lake Analytics/U-SQL

demo

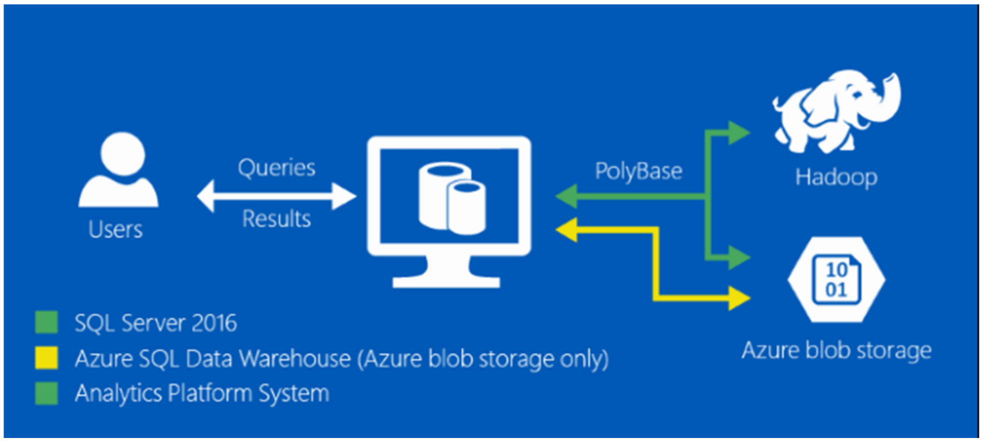

Users
Queries
Results
PolyBase
Hadoop
Azure blob storage
SQL Server 2016
Azure SQL Data Warehouse (Azure blob storage only)
Analytics Platform System

POLYBASE

PolyBase

- A “bridging” technology to connect SQL Server to data in Hadoop or Azure Blob Storage
- Makes the Hadoop data look like SQL Server data via “EXTERNAL” tables
- Query as normal; even join with physical tables
- First appeared in Parallel Data Warehouse/APS and Azure SQL DW
- Now included in SQL Server 2016 Enterprise
- Can create physical table with CREATE TABLE…AS SELECT… (CTAS)

PolyBase



Users

Queries

Results

PolyBase

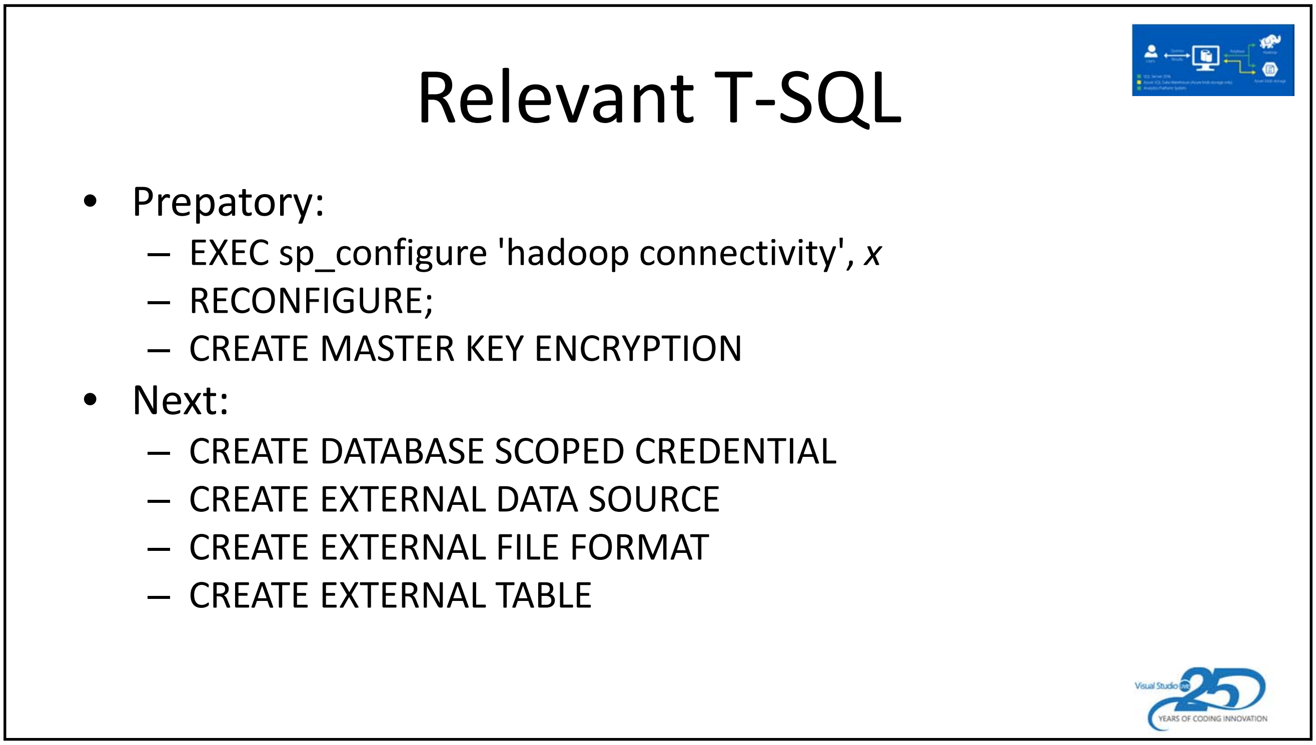
Hadoop

Azure blob storage

- 🟩 SQL Server 2016
- 🟨 Azure SQL Data Warehouse (Azure blob storage only)
- 🟩 Analytics Platform System

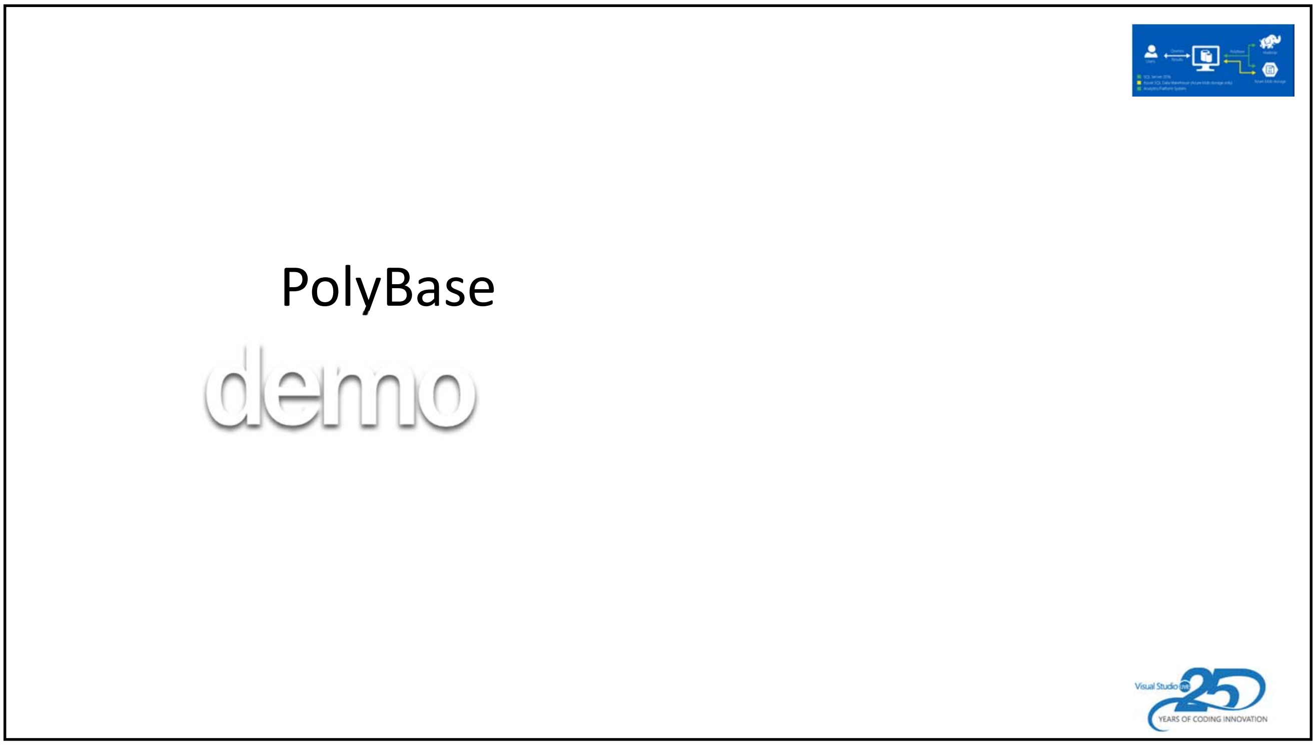
Notes

- Data may be moved and processed by SQL Server’s engine and optimizer, or may be “pushed down” to Hadoop, or both
- For DW versions of SQL, query is distributed
- Config can be tricky
- Java install is a prerequisite



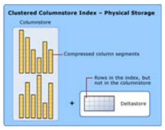
Relevant T-SQL

- Prepatory:
  - EXEC sp_configure 'hadoop connectivity', *x*
  - RECONFIGURE;
  - CREATE MASTER KEY ENCRYPTION
- Next:
  - CREATE DATABASE SCOPED CREDENTIAL
  - CREATE EXTERNAL DATA SOURCE
  - CREATE EXTERNAL FILE FORMAT
  - CREATE EXTERNAL TABLE

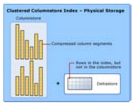


PolyBase

demo

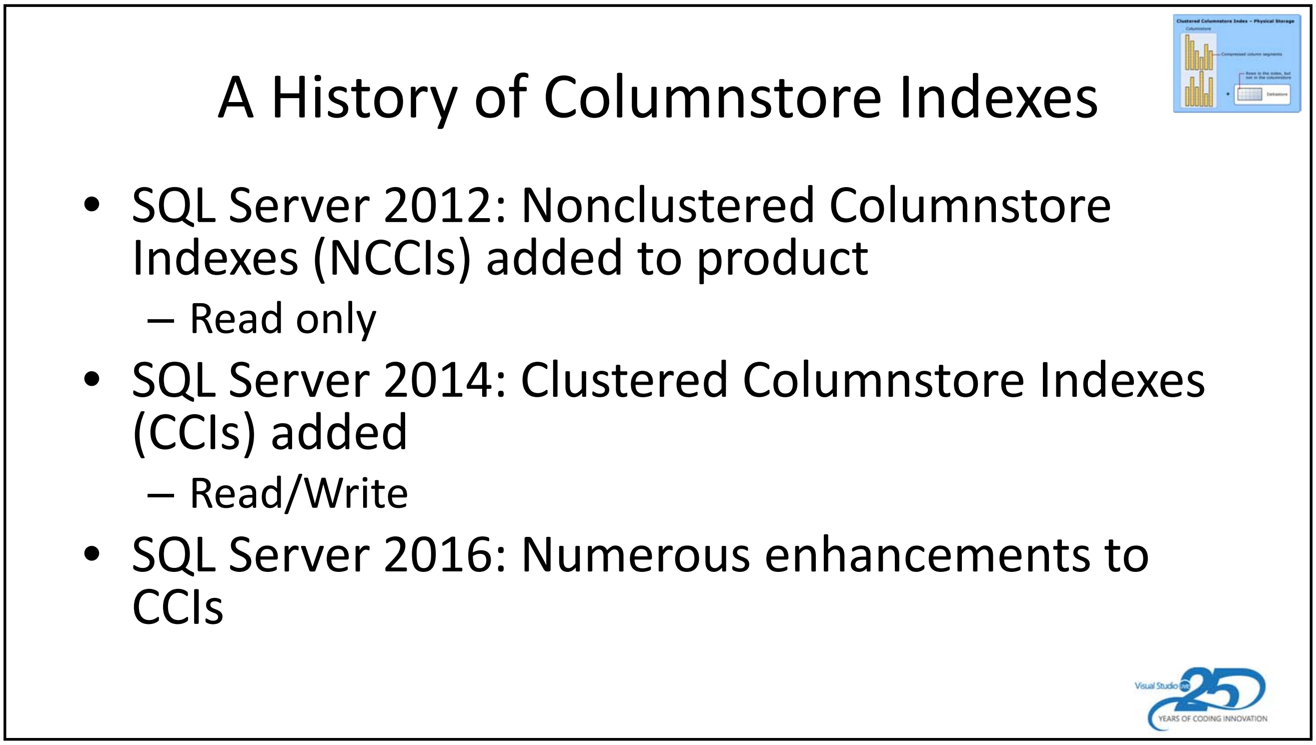

Clustered Columnstore Index – Physical Storage
Columnstore
Compressed column segments
Rows in the index, but not in the columnstore
+
Deltastore

COLUMNSTORE INDEXES


Clustered Columnstore Index – Physical Storage
Columnstore
Compressed column segments
Rows in the index, but not in the columnstore
+
Deltastore

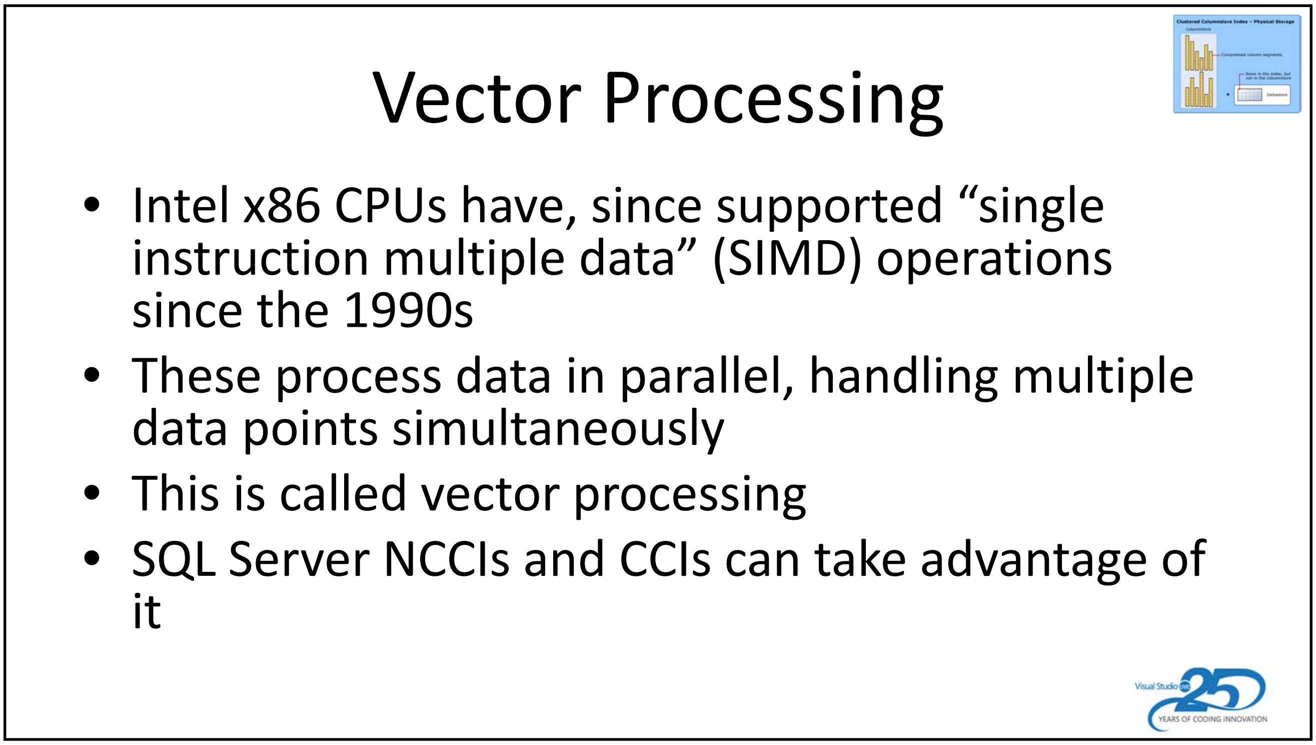
In Analytics…

- Geared to reporting and visualization
  - Read frequently, write seldom
- Table scans are expected
- Aggregation (think GROUP BY) is *de riguer*
- Extensive normalization is bad
- You only care about values in a small set of columns…maybe even just one
  - The rest are used with WHERE and HAVING, to filter
- Tables that track location and time are common



A History of Columnstore Indexes

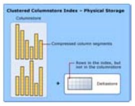
- SQL Server 2012: Nonclustered Columnstore Indexes (NCCIs) added to product
  - Read only
- SQL Server 2014: Clustered Columnstore Indexes (CCIs) added
  - Read/Write
- SQL Server 2016: Numerous enhancements to CCIs



Vector Processing

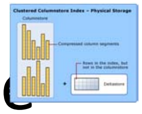
- Intel x86 CPUs have, since supported “single instruction multiple data” (SIMD) operations since the 1990s
- These process data in parallel, handling multiple data points simultaneously
- This is called vector processing
- SQL Server NCCIs and CCIs can take advantage of it

Useful Applications


Clustered Columnstore Index – Physical Storage
Columnstore
Compressed column segments
Rows in the index, but not in the columnstore
+
Deltastore

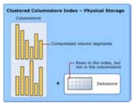
- Data Warehouse/Data Mart scenarios
- In combination with DirectQuery feature in SSAS Tabular and Power BI
- In combination with R Services

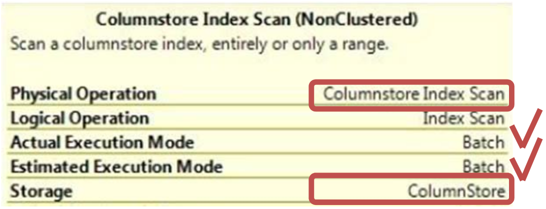
Vector Processing and “Batch” Mode


Clustered Columnstore Index – Physical Storage
Columnstore
Compressed column segments
Rows in the index, but not in the columnstore
+
Deltastore

- SQL Server can burst into a vector processing mode
- Instead of iterating through rowsets, one row at a time, it can handle rows in batches
- So it’s called “batch mode” and it’s *fast*
  - (Not to be confused with batch processing, which can be slow)

Sanity Check



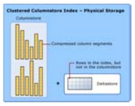


Columnstore Index Scan (NonClustered)

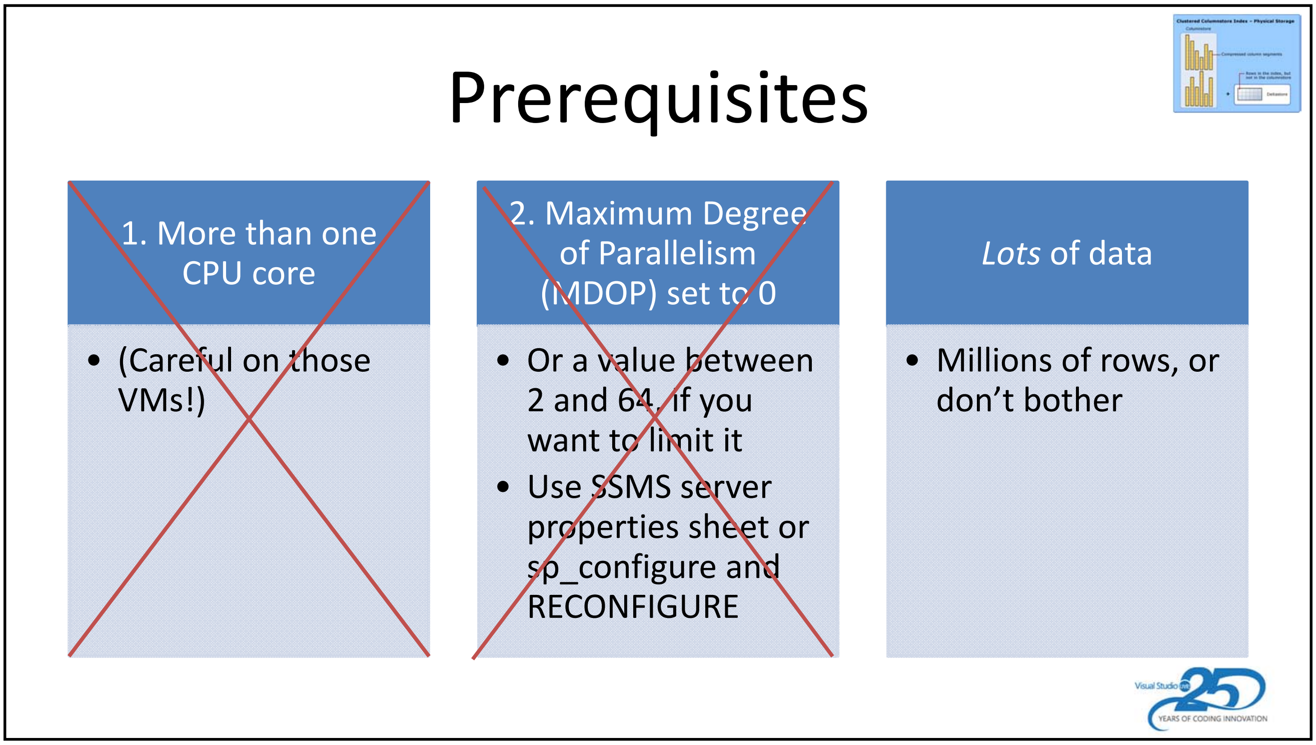
Scan a columnstore index, entirely or only a range.

Physical Operation	Columnstore Index Scan
Logical Operation	Index Scan
Actual Execution Mode	Batch ✓
Estimated Execution Mode	Batch ✓
Storage	ColumnStore

Making Sure it Works



- CIs are fastest when Batch mode kicks in
  - Difference can be negligible otherwise
  - Check Query Plan to make sure
- And meet the prerequisites…



Prerequisites

~~1. More than one CPU core~~

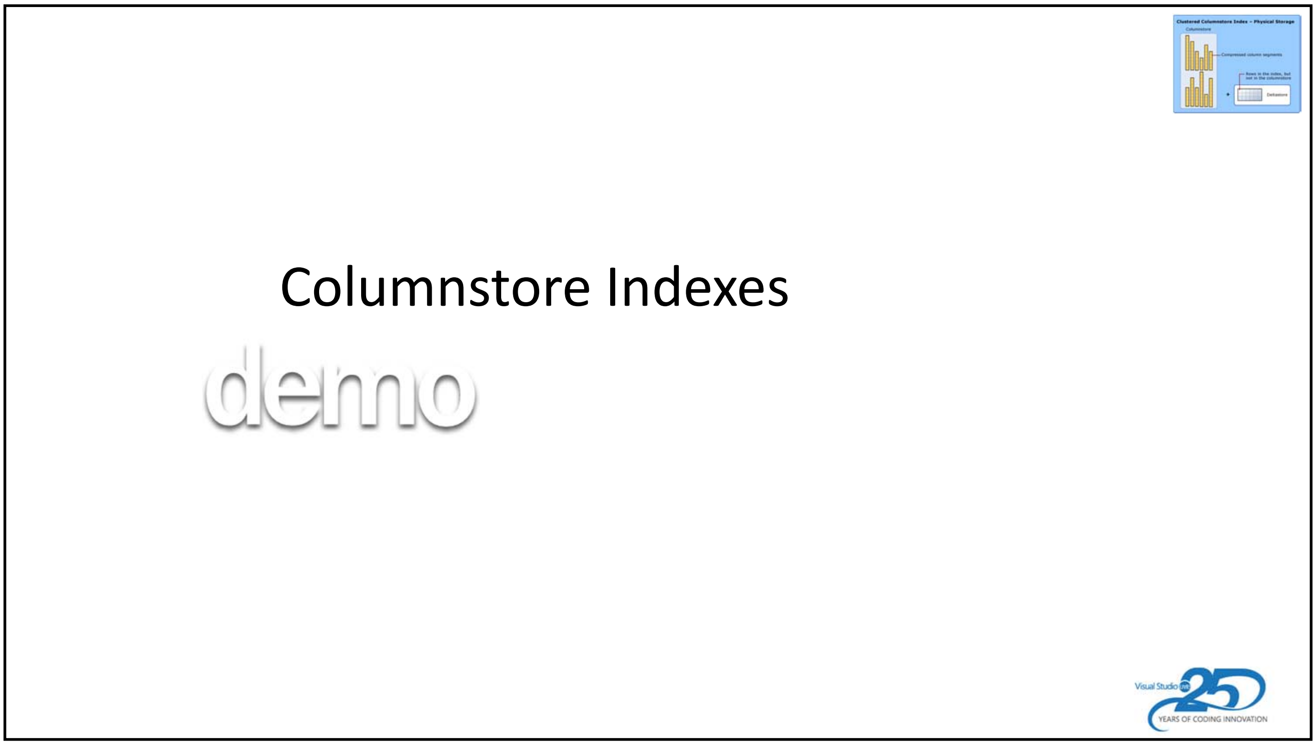
- ~~(Careful on those VMs!)~~

~~2. Maximum Degree of Parallelism (MDOP) set to 0~~

- ~~Or a value between 2 and 64, if you want to limit it~~
- ~~Use SSMS server properties sheet or sp_configure and RECONFIGURE~~

Lots of data

- Millions of rows, or don't bother



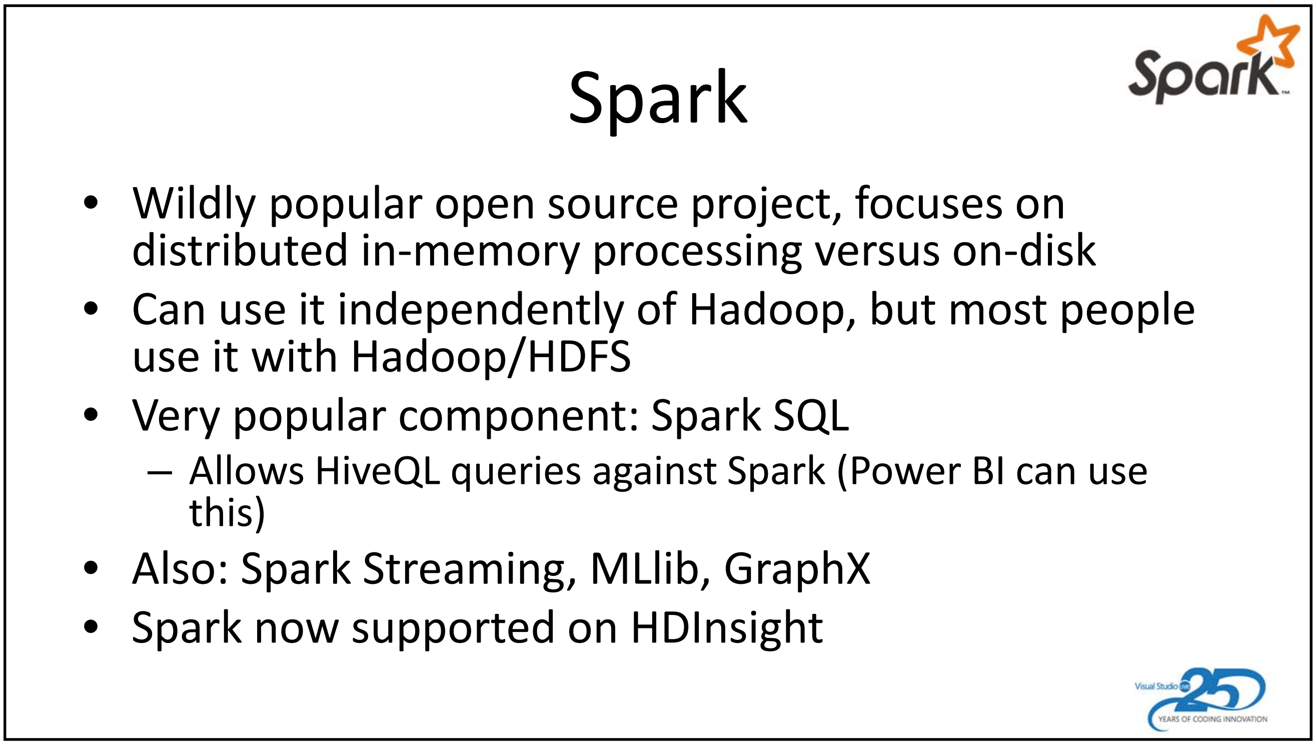
Columnstore Indexes

demo

Distributed, In-Memory

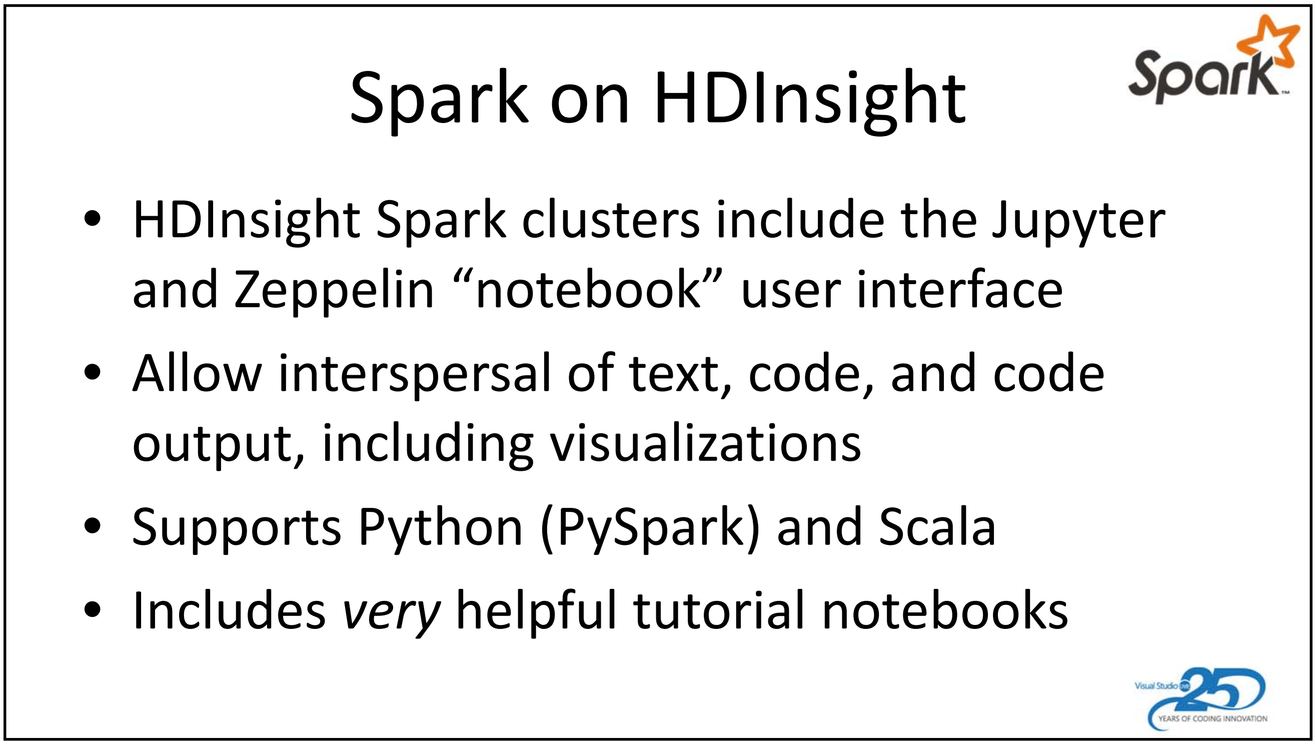
Big Data Platform

APACHE SPARK



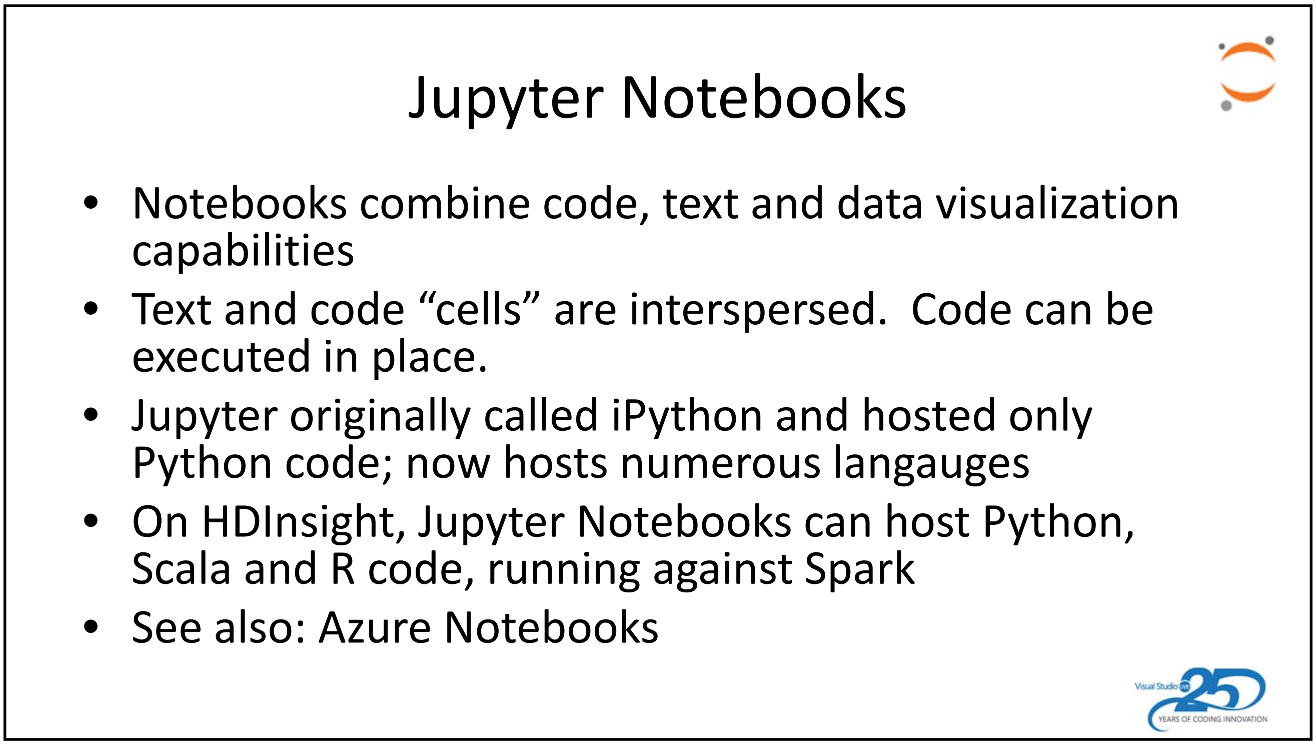
Spark

- Wildly popular open source project, focuses on distributed in-memory processing versus on-disk
- Can use it independently of Hadoop, but most people use it with Hadoop/HDFS
- Very popular component: Spark SQL
  - Allows HiveQL queries against Spark (Power BI can use this)
- Also: Spark Streaming, MLlib, GraphX
- Spark now supported on HDInsight



Spark on HDInsight

- HDInsight Spark clusters include the Jupyter and Zeppelin “notebook” user interface
- Allow interspersal of text, code, and code output, including visualizations
- Supports Python (PySpark) and Scala
- Includes *very* helpful tutorial notebooks



Jupyter Notebooks

- Notebooks combine code, text and data visualization capabilities
- Text and code “cells” are interspersed. Code can be executed in place.
- Jupyter originally called iPython and hosted only Python code; now hosts numerous langauges
- On HDInsight, Jupyter Notebooks can host Python, Scala and R code, running against Spark
- See also: Azure Notebooks

Apache Spark

demo

The Hadoop Stack

Security, governance

Stream processing, analytics

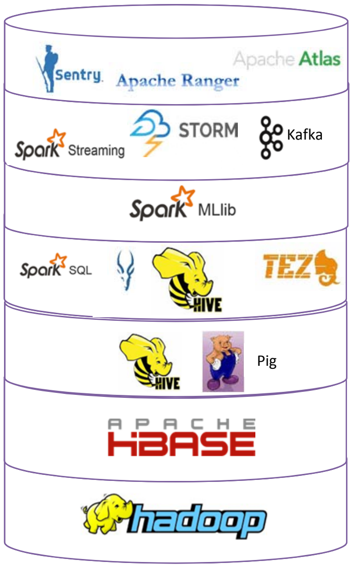
Machine Learning

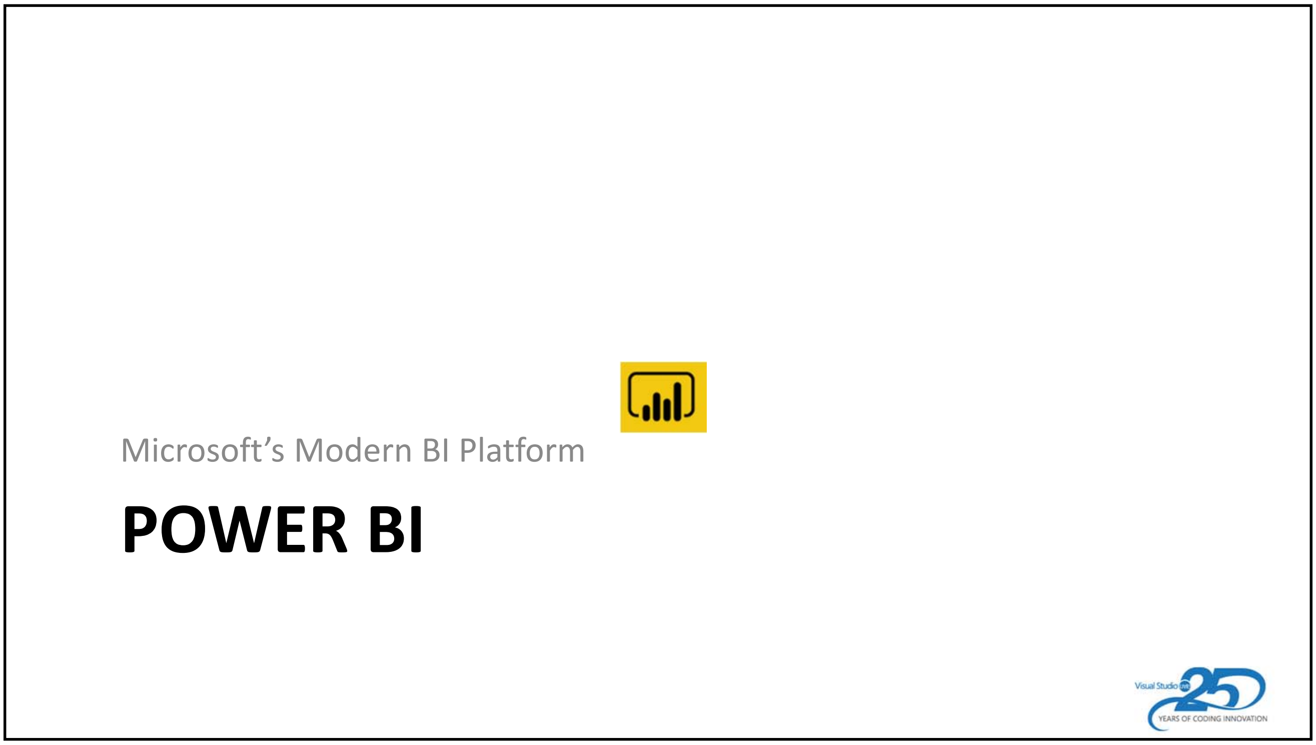
Interactive SQL

Query: HiveQL and Pig Latin

Database (NoSQL)

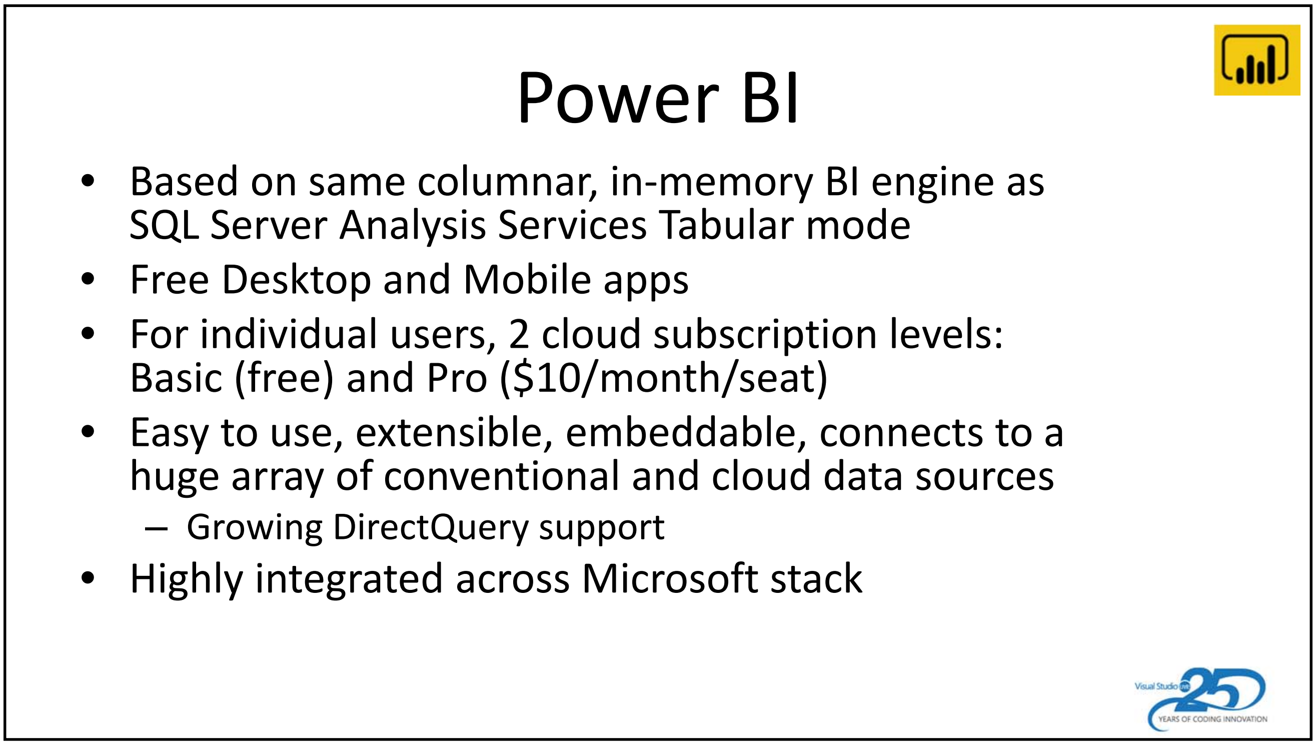
HDFS, YARN





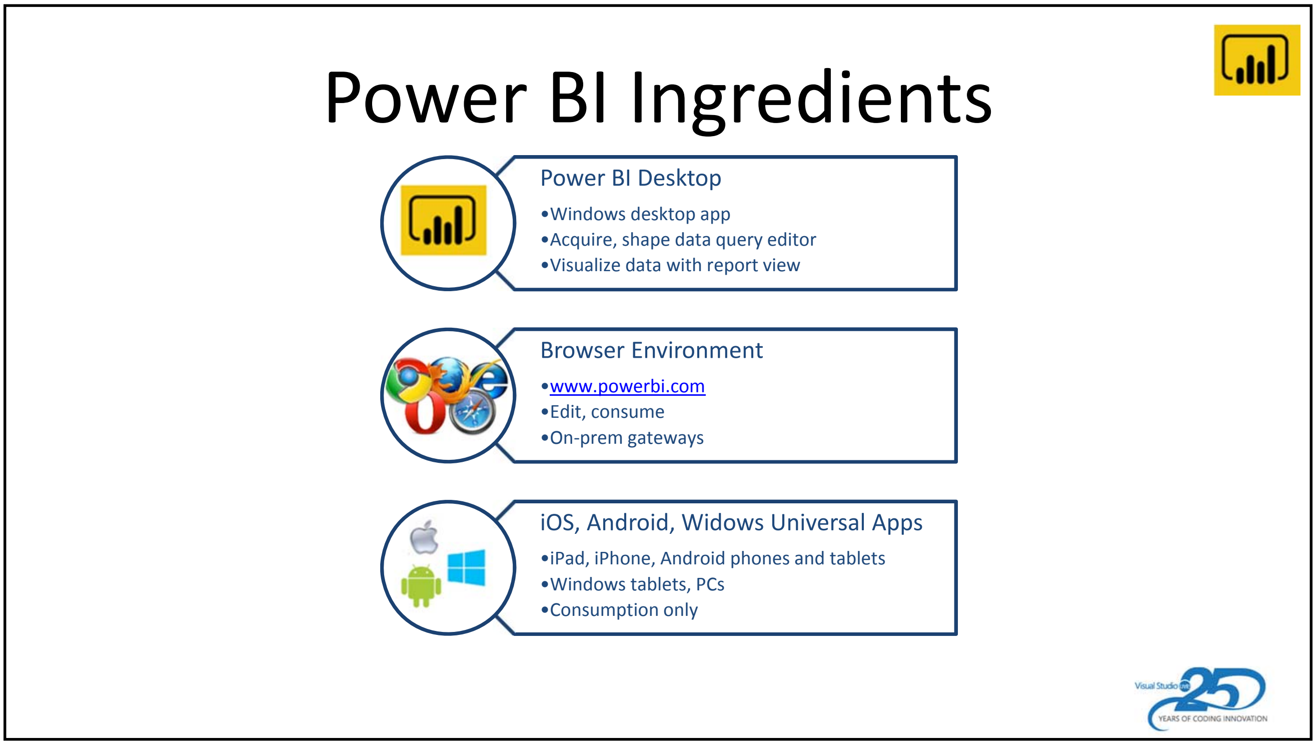
Microsoft’s Modern BI Platform

POWER BI



Power BI

- Based on same columnar, in-memory BI engine as SQL Server Analysis Services Tabular mode
- Free Desktop and Mobile apps
- For individual users, 2 cloud subscription levels: Basic (free) and Pro ($10/month/seat)
- Easy to use, extensible, embeddable, connects to a huge array of conventional and cloud data sources
  - Growing DirectQuery support
- Highly integrated across Microsoft stack

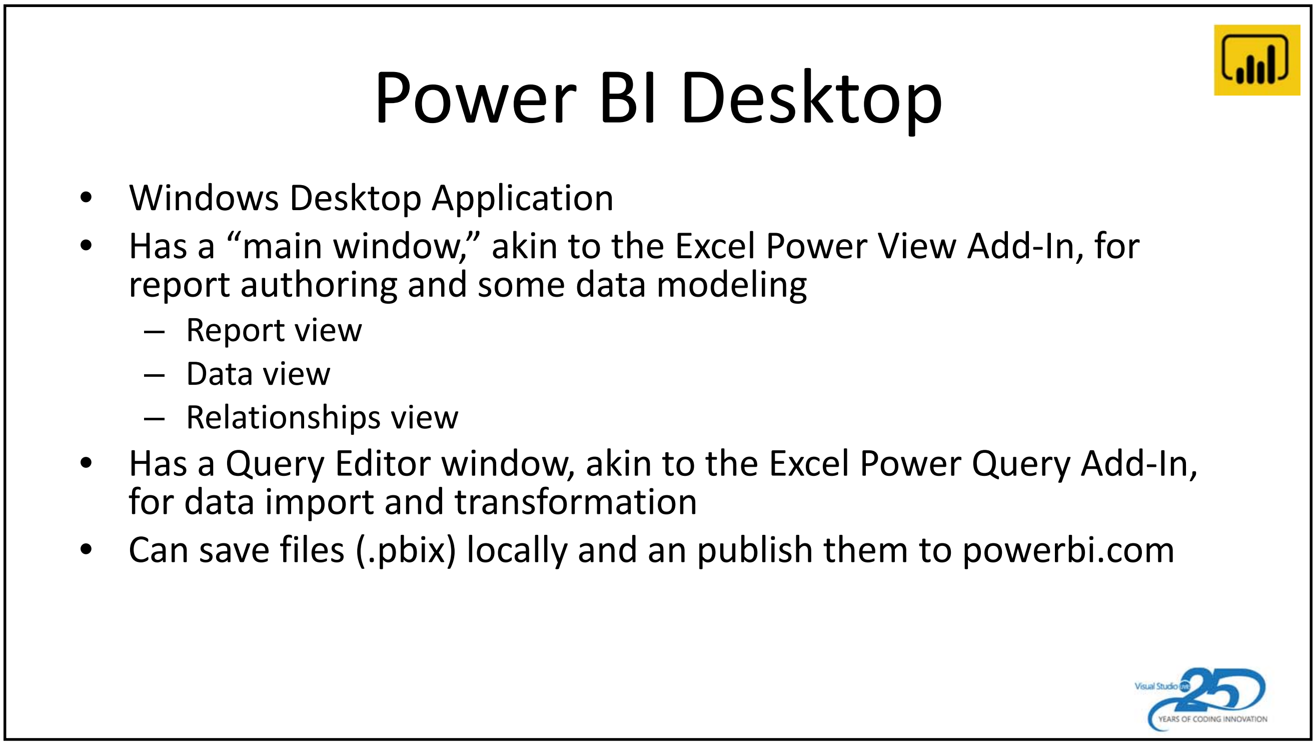


Power BI Ingredients

Power BI Desktop
- Windows desktop app
- Acquire, shape data query editor
- Visualize data with report view

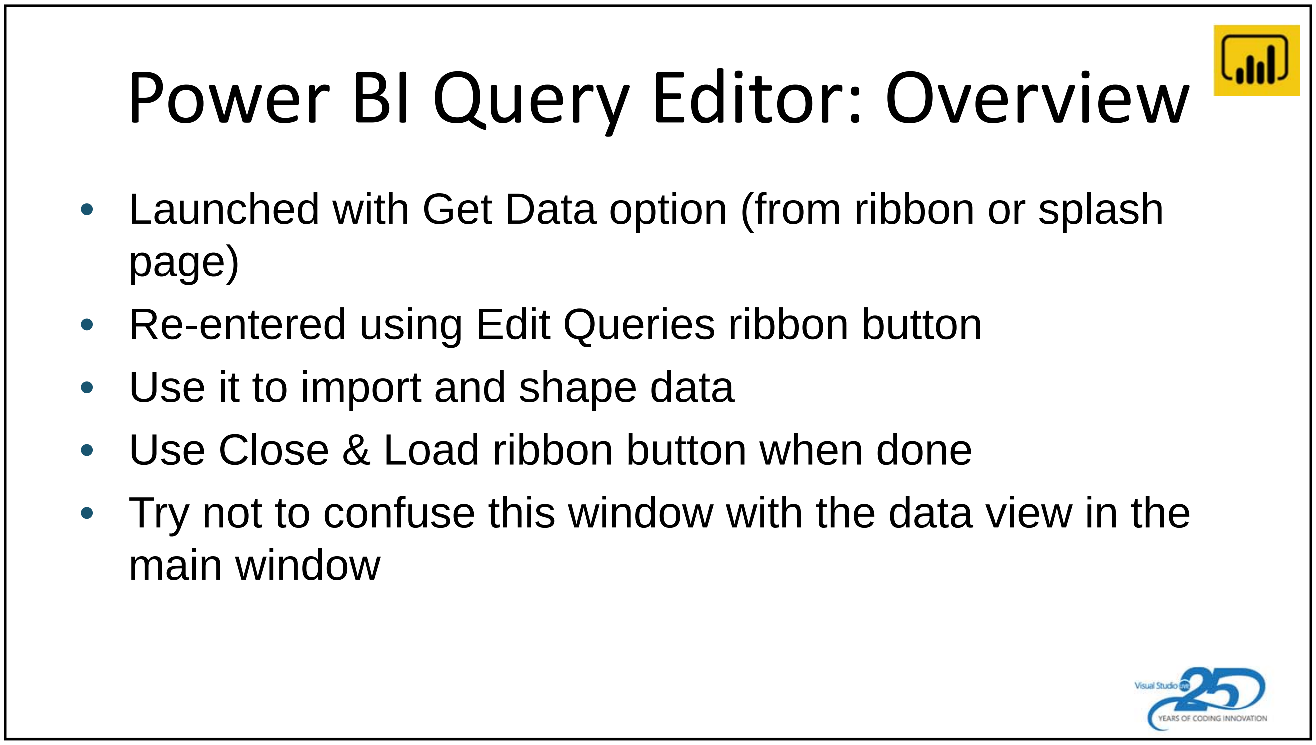
Browser Environment
- www.powerbi.com
- Edit, consume
- On-prem gateways

iOS, Android, Widows Universal Apps
- iPad, iPhone, Android phones and tablets
- Windows tablets, PCs
- Consumption only



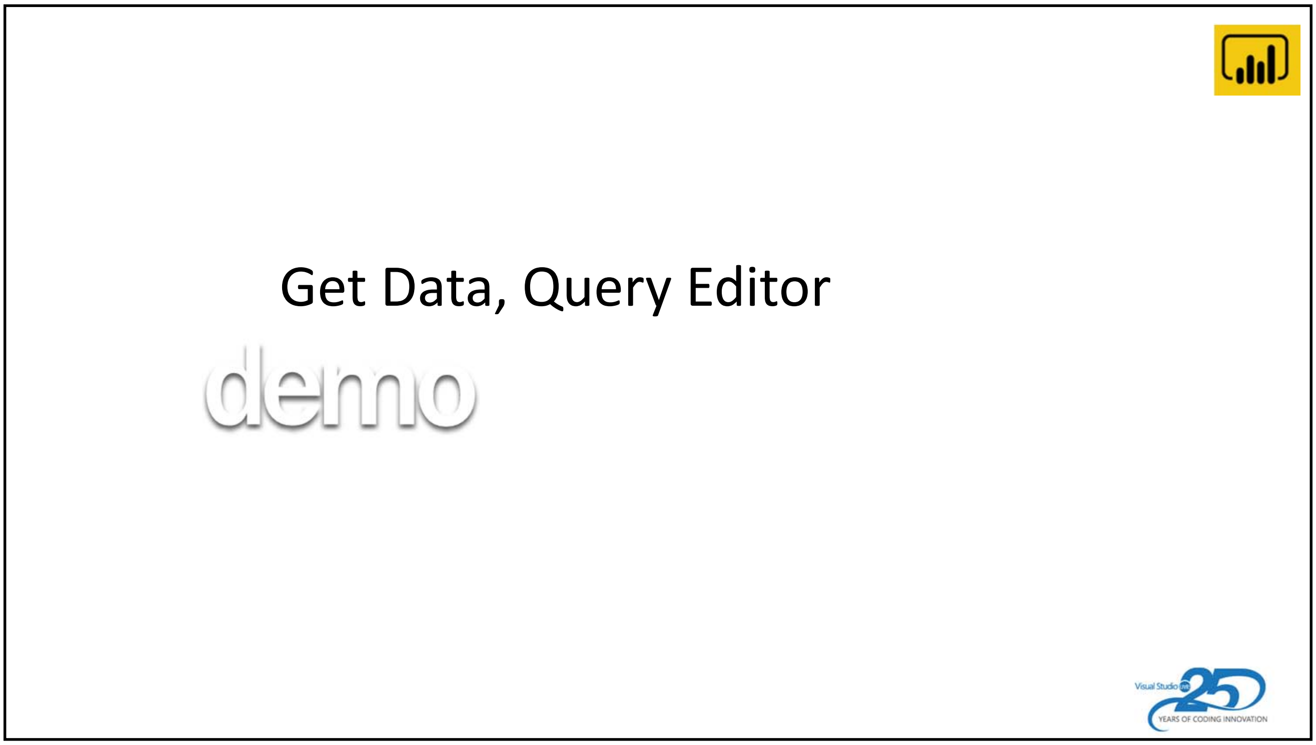
Power BI Desktop

- Windows Desktop Application
- Has a “main window,” akin to the Excel Power View Add-In, for report authoring and some data modeling
  - Report view
  - Data view
  - Relationships view
- Has a Query Editor window, akin to the Excel Power Query Add-In, for data import and transformation
- Can save files (.pbix) locally and an publish them to powerbi.com



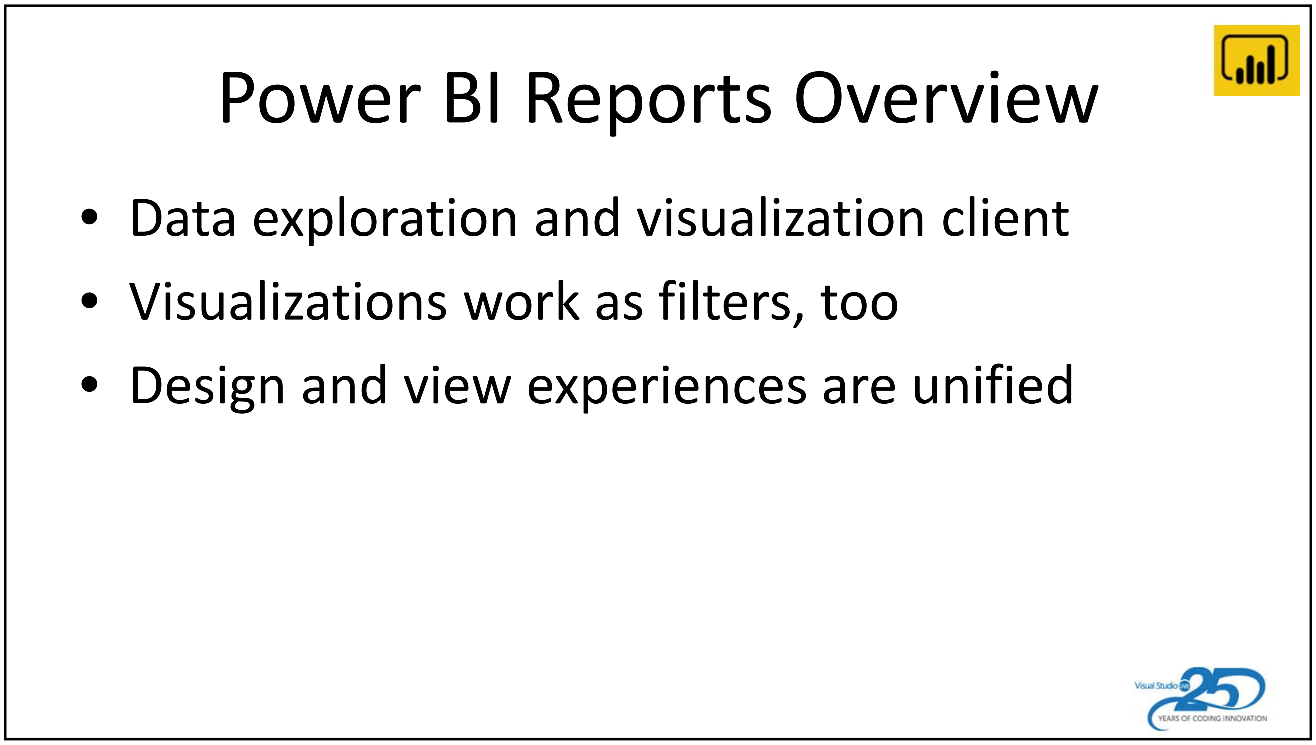
Power BI Query Editor: Overview

- Launched with Get Data option (from ribbon or splash page)
- Re-entered using Edit Queries ribbon button
- Use it to import and shape data
- Use Close & Load ribbon button when done
- Try not to confuse this window with the data view in the main window



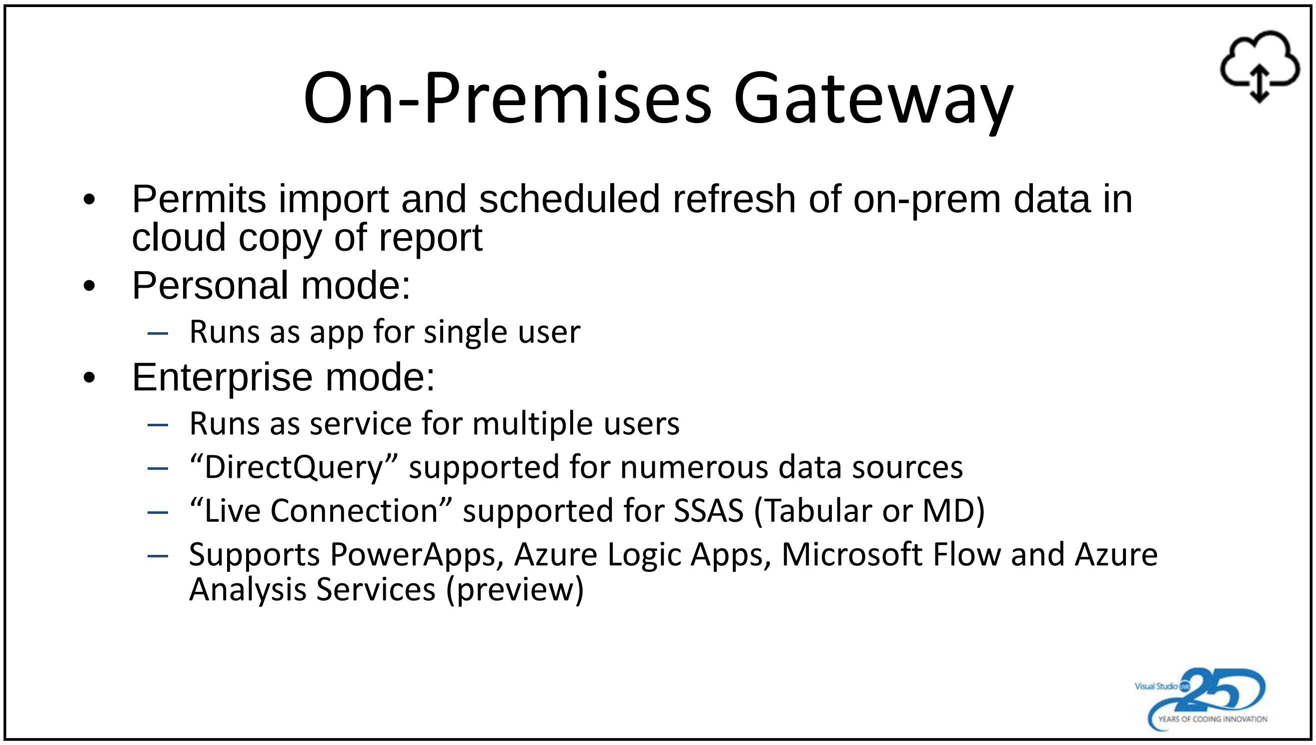
Get Data, Query Editor

demo



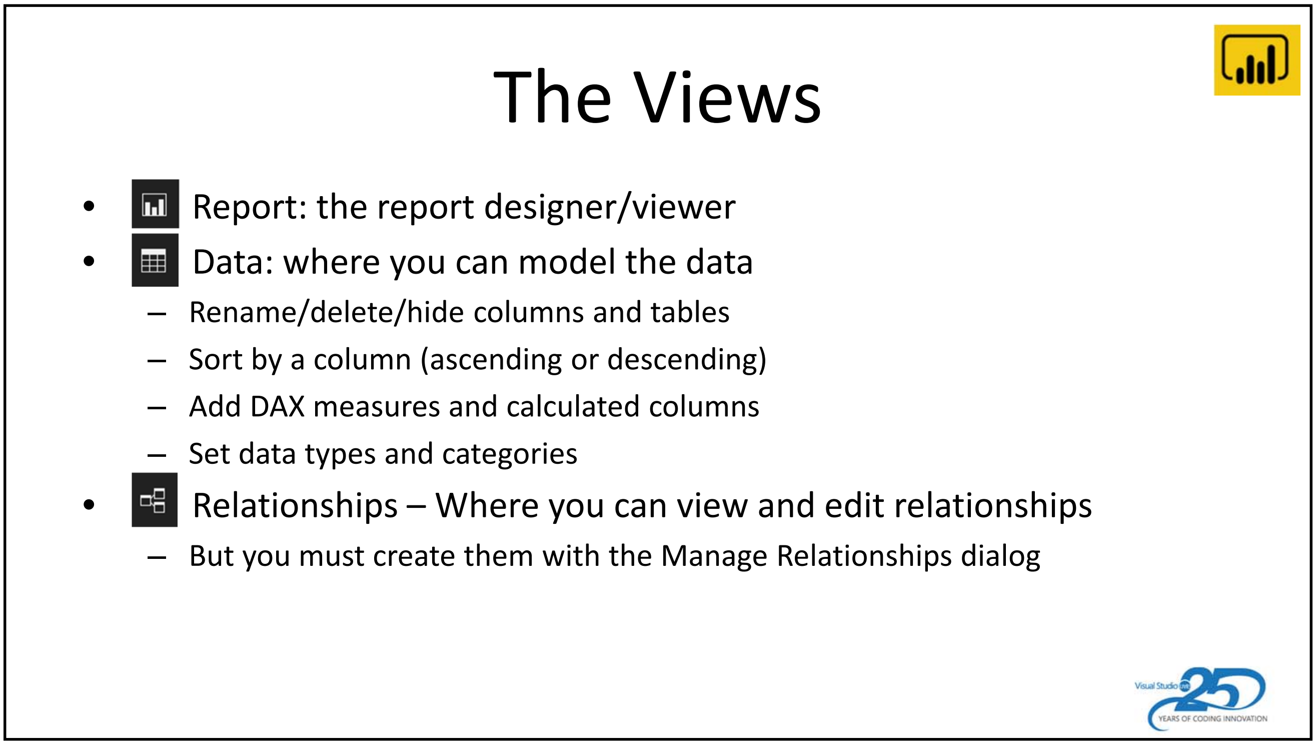
Power BI Reports Overview

- Data exploration and visualization client
- Visualizations work as filters, too
- Design and view experiences are unified



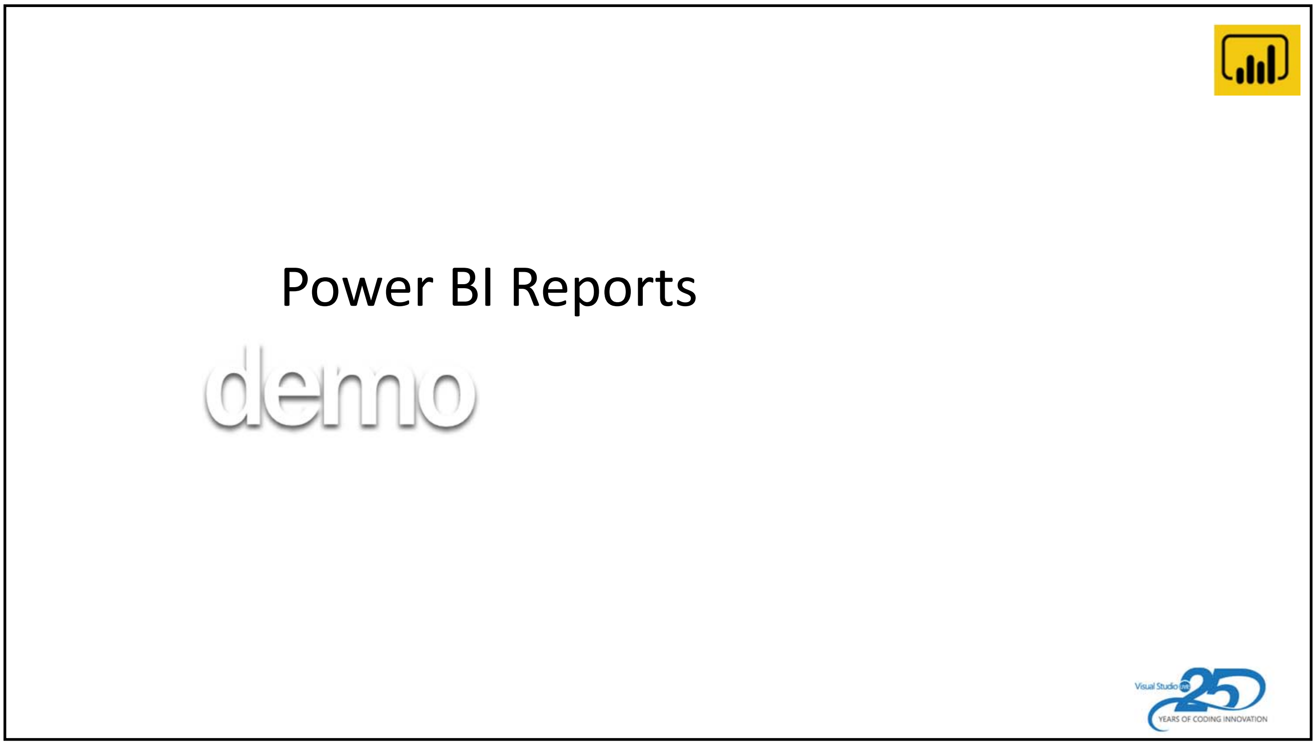
On-Premises Gateway

- Permits import and scheduled refresh of on-prem data in cloud copy of report
- Personal mode:
  - Runs as app for single user
- Enterprise mode:
  - Runs as service for multiple users
  - "DirectQuery" supported for numerous data sources
  - "Live Connection" supported for SSAS (Tabular or MD)
  - Supports PowerApps, Azure Logic Apps, Microsoft Flow and Azure Analysis Services (preview)



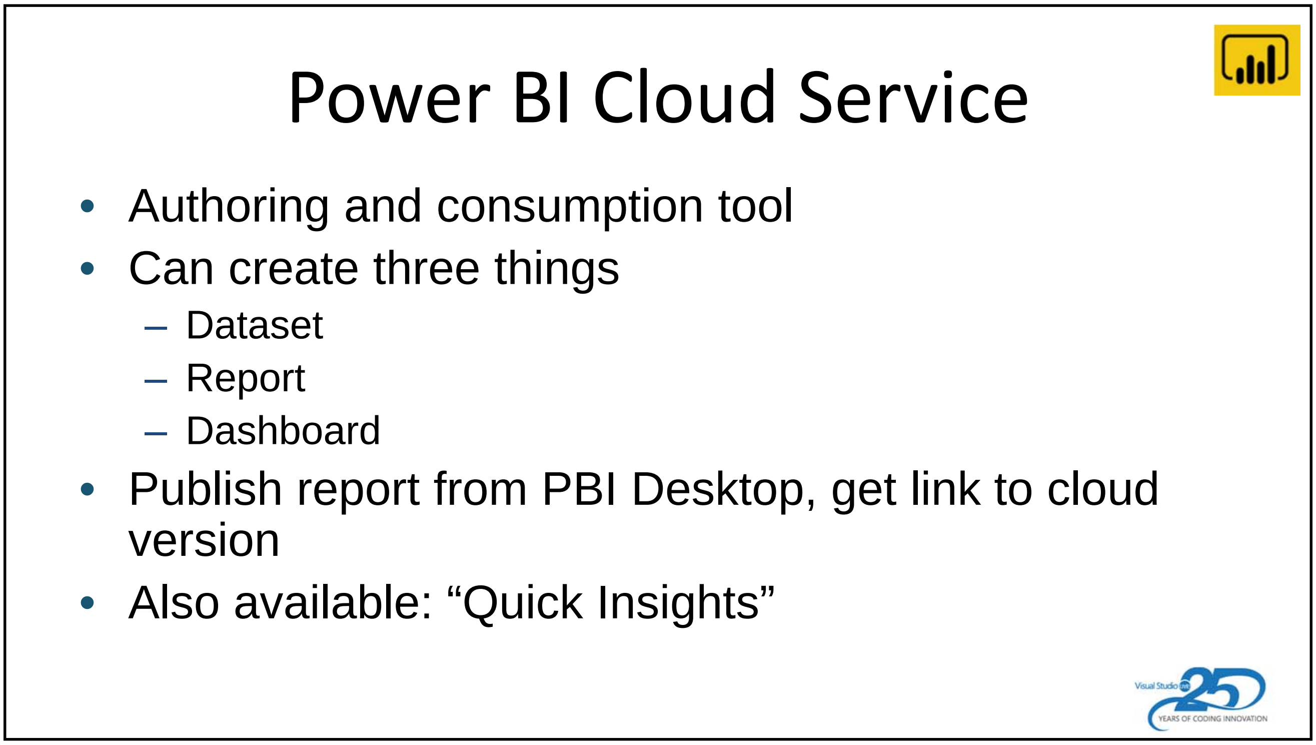
The Views

- Report: the report designer/viewer
- Data: where you can model the data
  - Rename/delete/hide columns and tables
  - Sort by a column (ascending or descending)
  - Add DAX measures and calculated columns
  - Set data types and categories
- Relationships – Where you can view and edit relationships
  - But you must create them with the Manage Relationships dialog



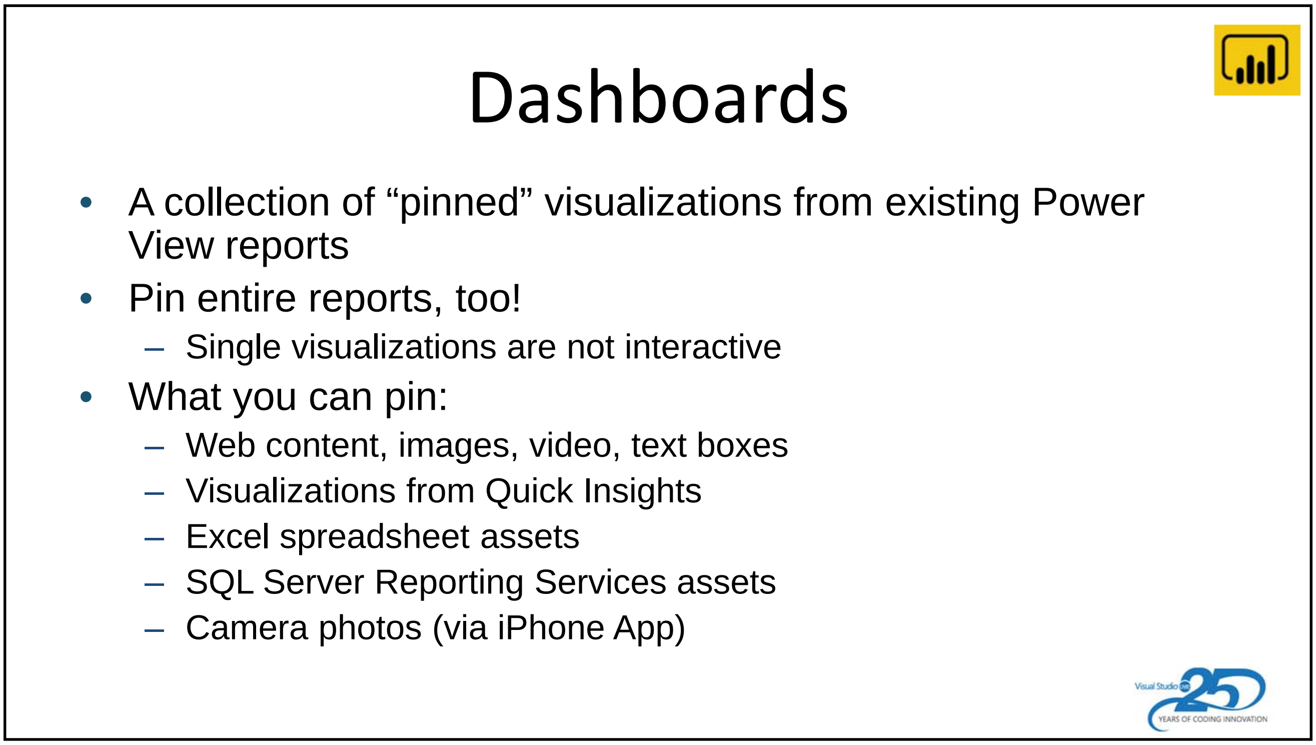
Power BI Reports

demo



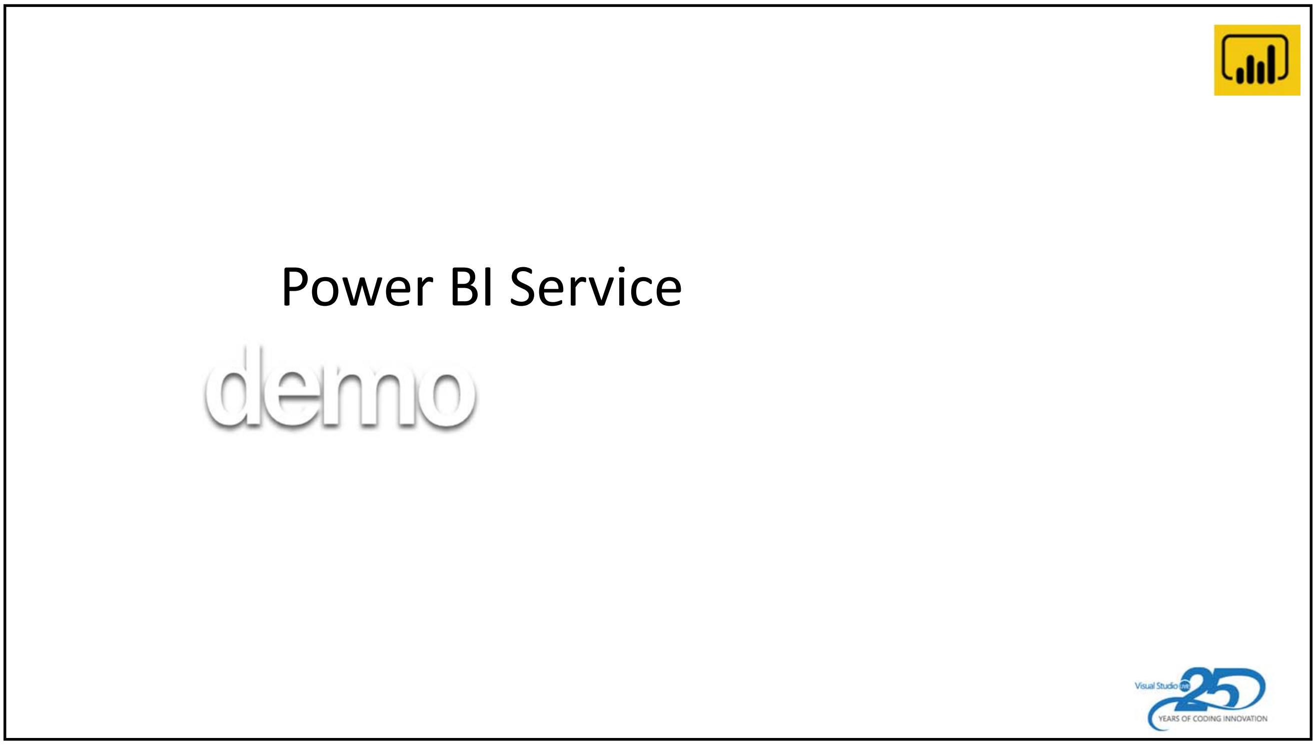
Power BI Cloud Service

- Authoring and consumption tool
- Can create three things
  - Dataset
  - Report
  - Dashboard
- Publish report from PBI Desktop, get link to cloud version
- Also available: “Quick Insights”



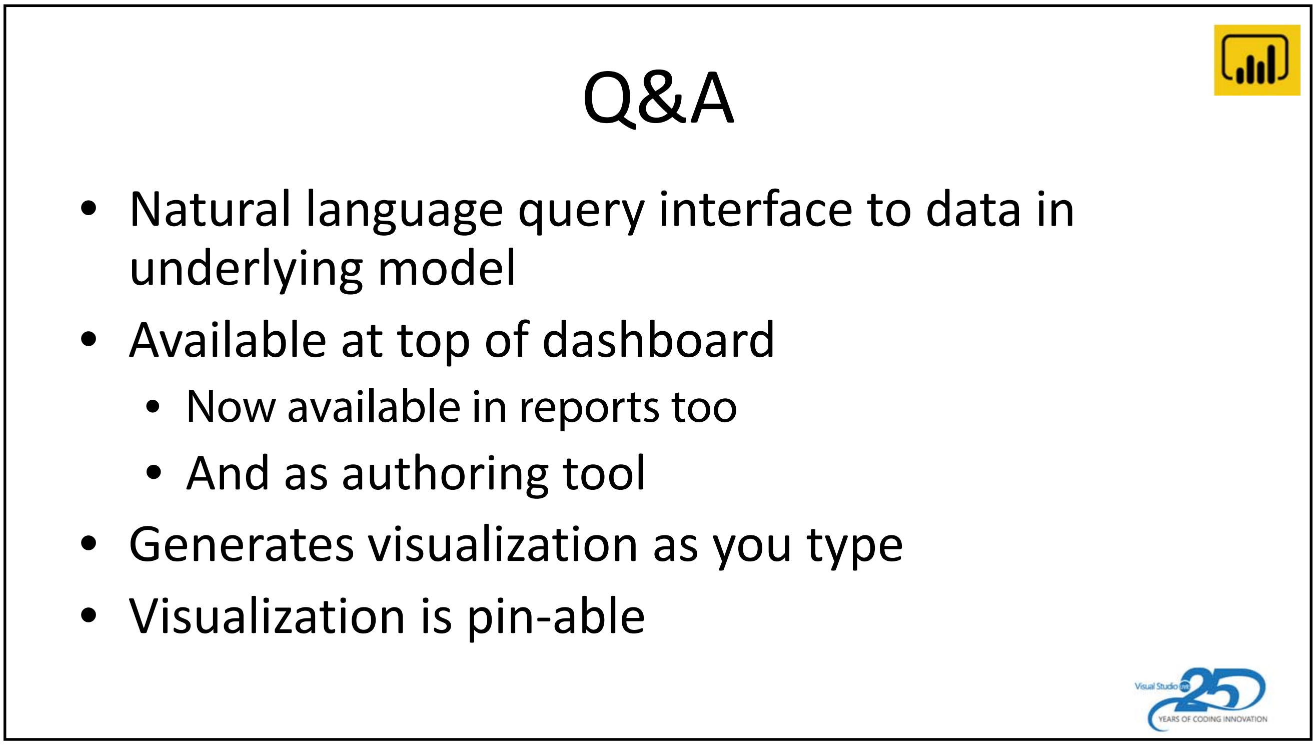
Dashboards

- A collection of “pinned” visualizations from existing Power View reports
- Pin entire reports, too!
  - Single visualizations are not interactive
- What you can pin:
  - Web content, images, video, text boxes
  - Visualizations from Quick Insights
  - Excel spreadsheet assets
  - SQL Server Reporting Services assets
  - Camera photos (via iPhone App)



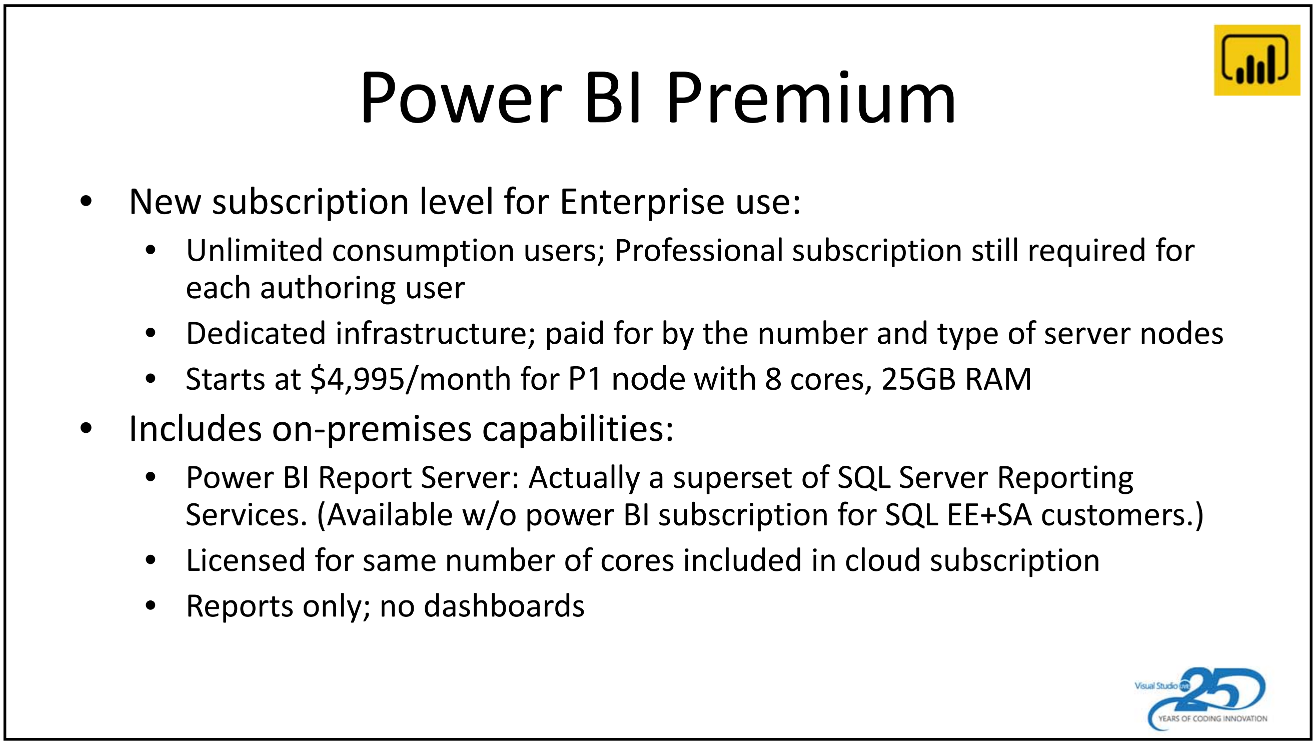
Power BI Service

demo



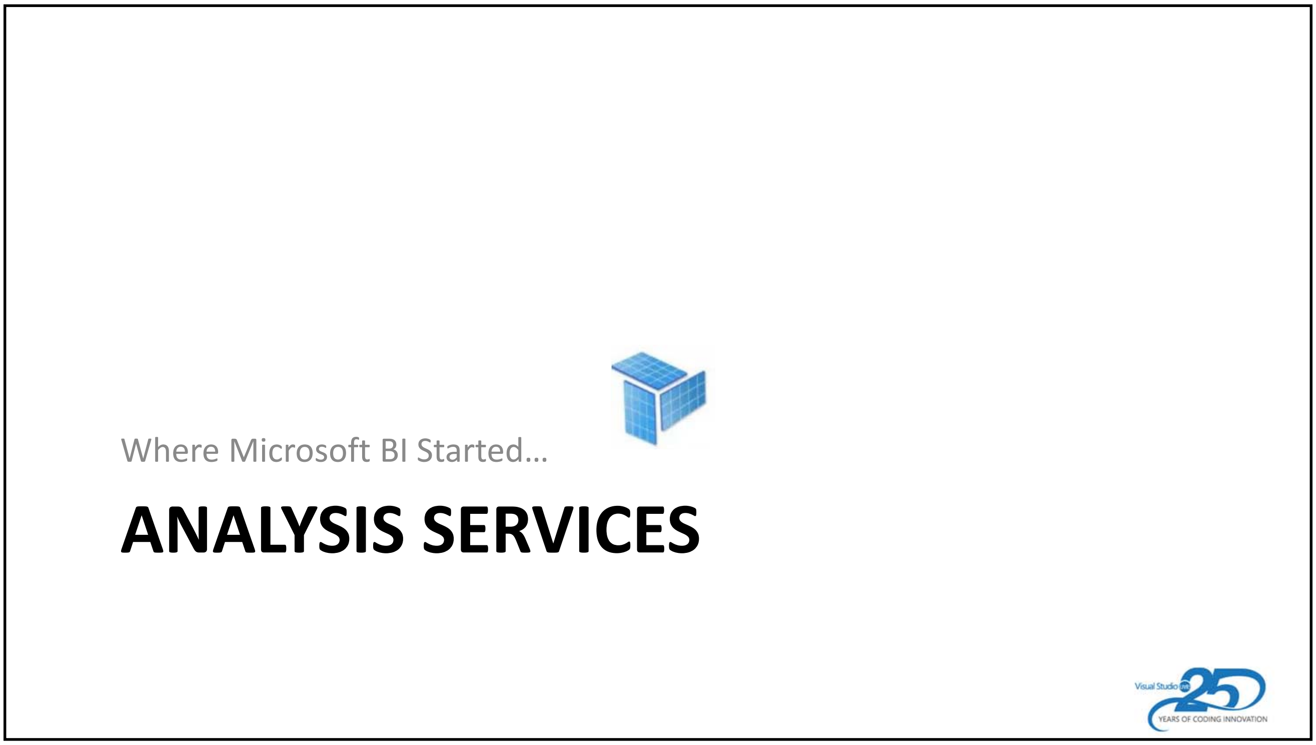
Q&A

- Natural language query interface to data in underlying model
- Available at top of dashboard
  - Now available in reports too
  - And as authoring tool
- Generates visualization as you type
- Visualization is pin-able



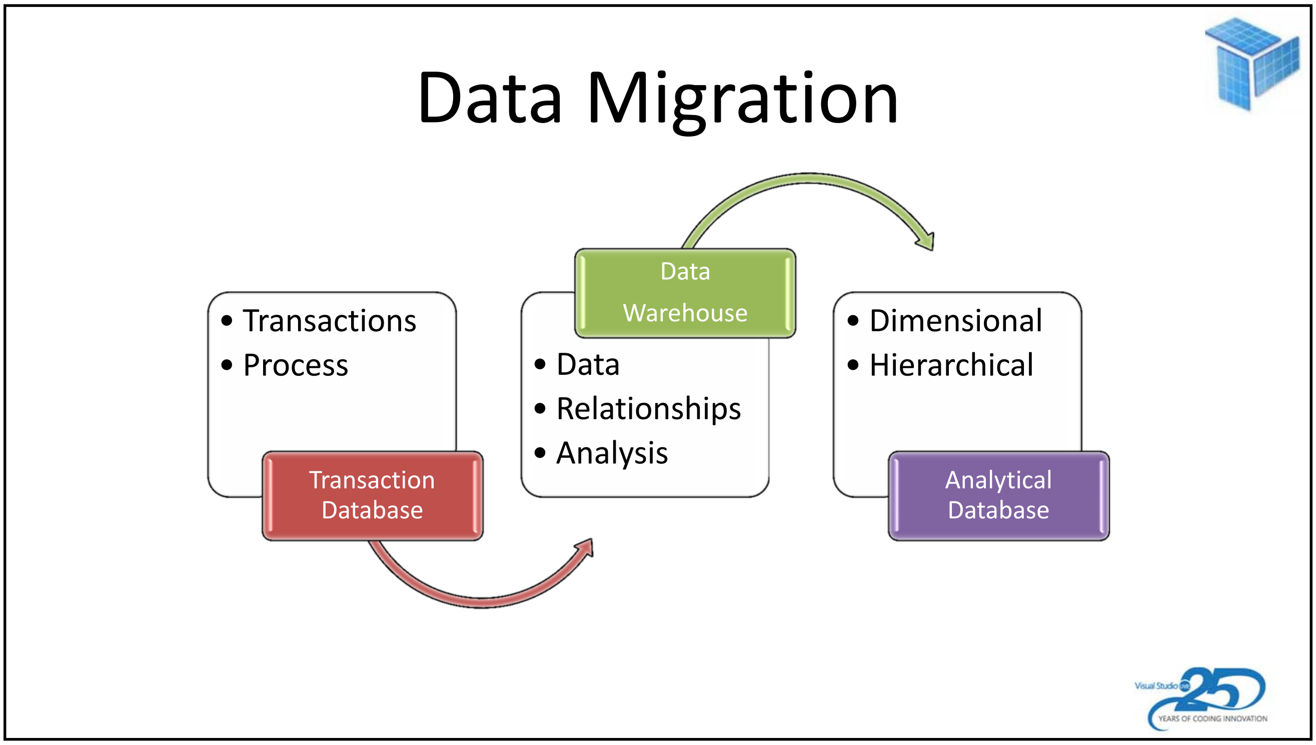
Power BI Premium

- New subscription level for Enterprise use:
  - Unlimited consumption users; Professional subscription still required for each authoring user
  - Dedicated infrastructure; paid for by the number and type of server nodes
  - Starts at $4,995/month for P1 node with 8 cores, 25GB RAM
- Includes on-premises capabilities:
  - Power BI Report Server: Actually a superset of SQL Server Reporting Services. (Available w/o power BI subscription for SQL EE+SA customers.)
  - Licensed for same number of cores included in cloud subscription
  - Reports only; no dashboards



Where Microsoft BI Started…

ANALYSIS SERVICES



Data Migration

- Transactions
- Process

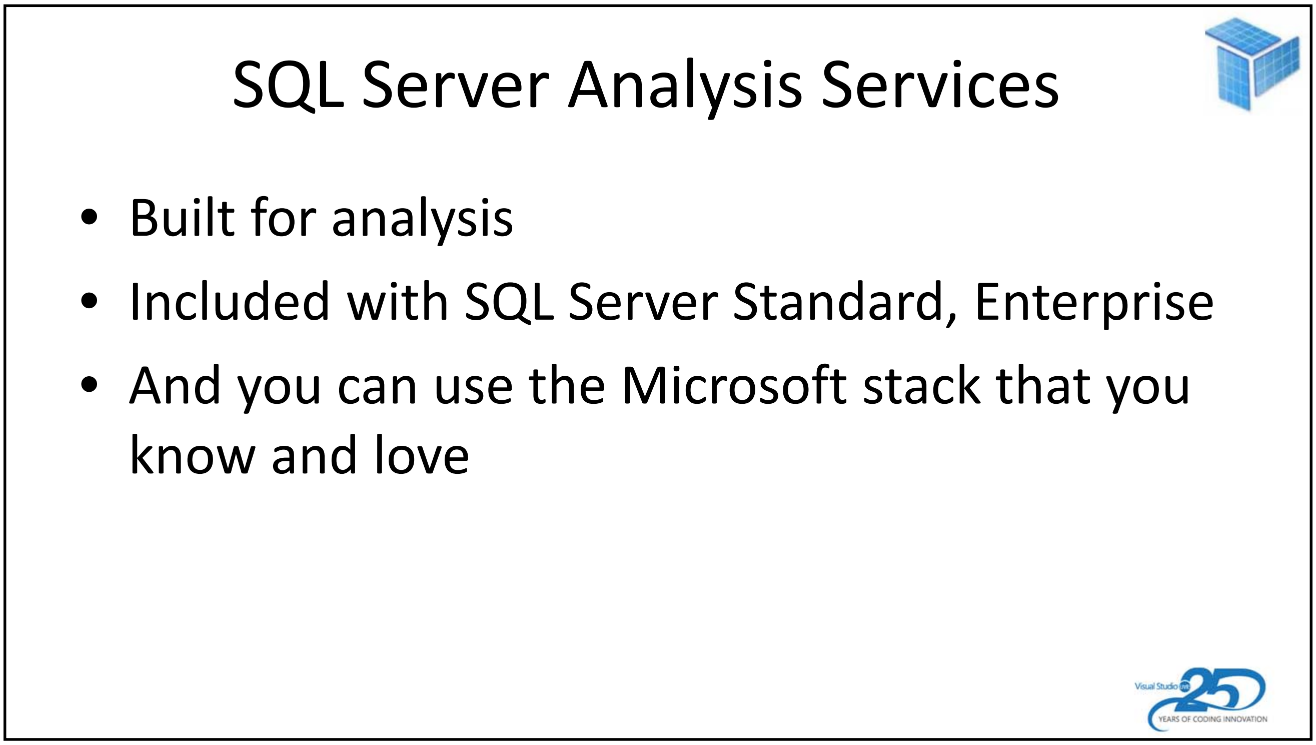
Transaction Database

Data Warehouse

- Data
- Relationships
- Analysis

- Dimensional
- Hierarchical

Analytical Database

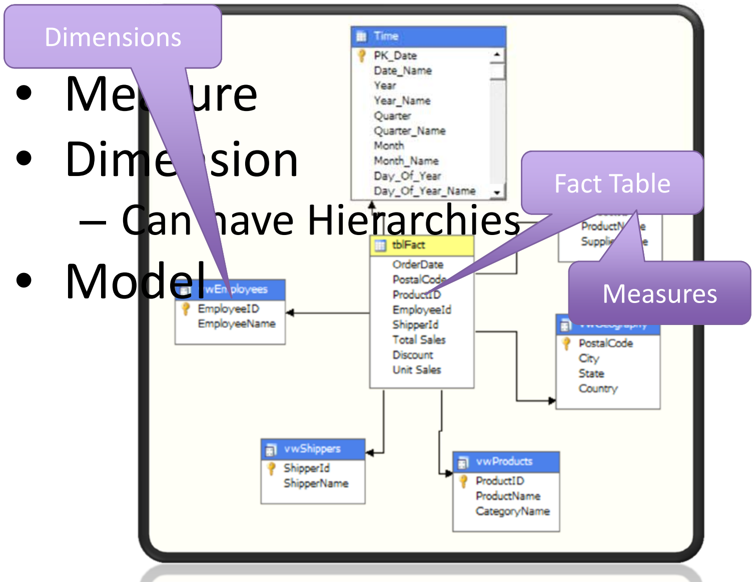


SQL Server Analysis Services

- Built for analysis
- Included with SQL Server Standard, Enterprise
- And you can use the Microsoft stack that you know and love

From Data Warehouse to OLAP

- Measure
- Dimension
  - Can have Hierarchies
- Model



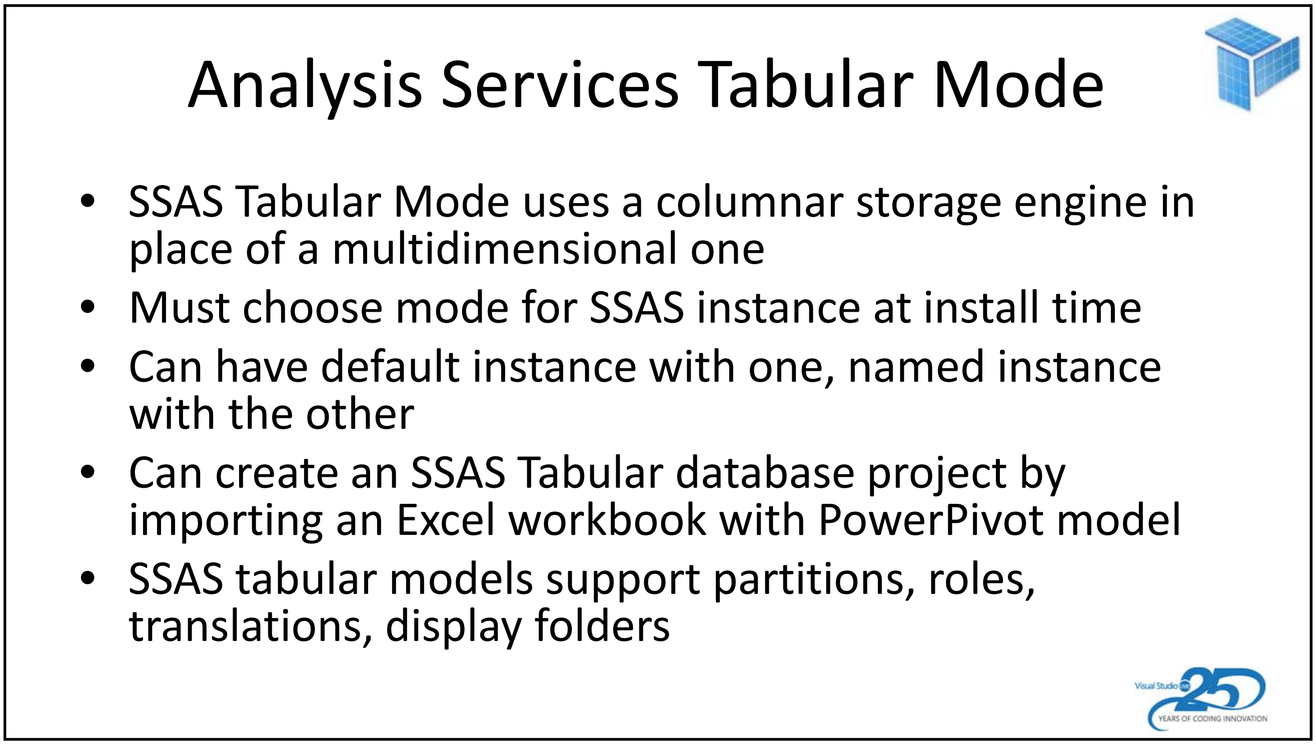
Dimensions

Fact Table

Measures

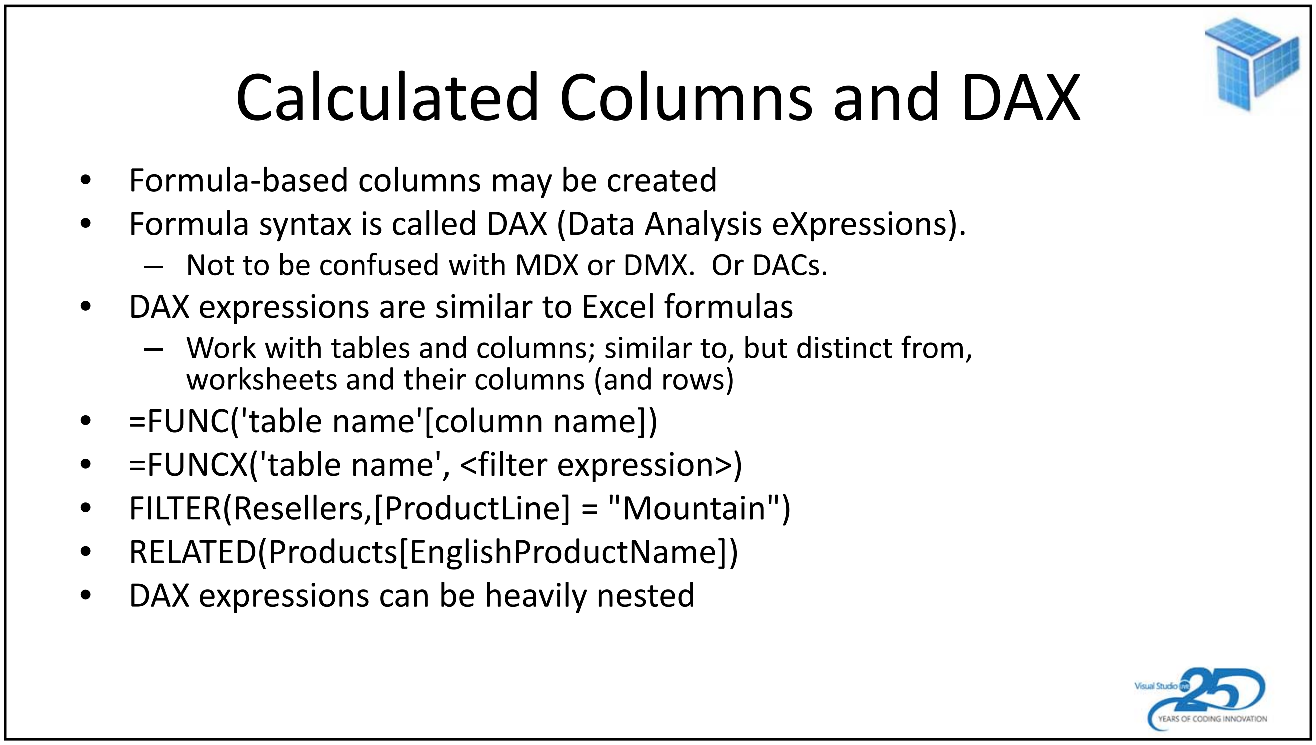
Analysis Services Modes

- **Multidimensional** or **Tabular**
- Tabular is newer, same tech as Excel/PowerPivot data models and Power BI
- Lots of investment in Tabular in SSAS 2016
- We’ll look at Tabular today



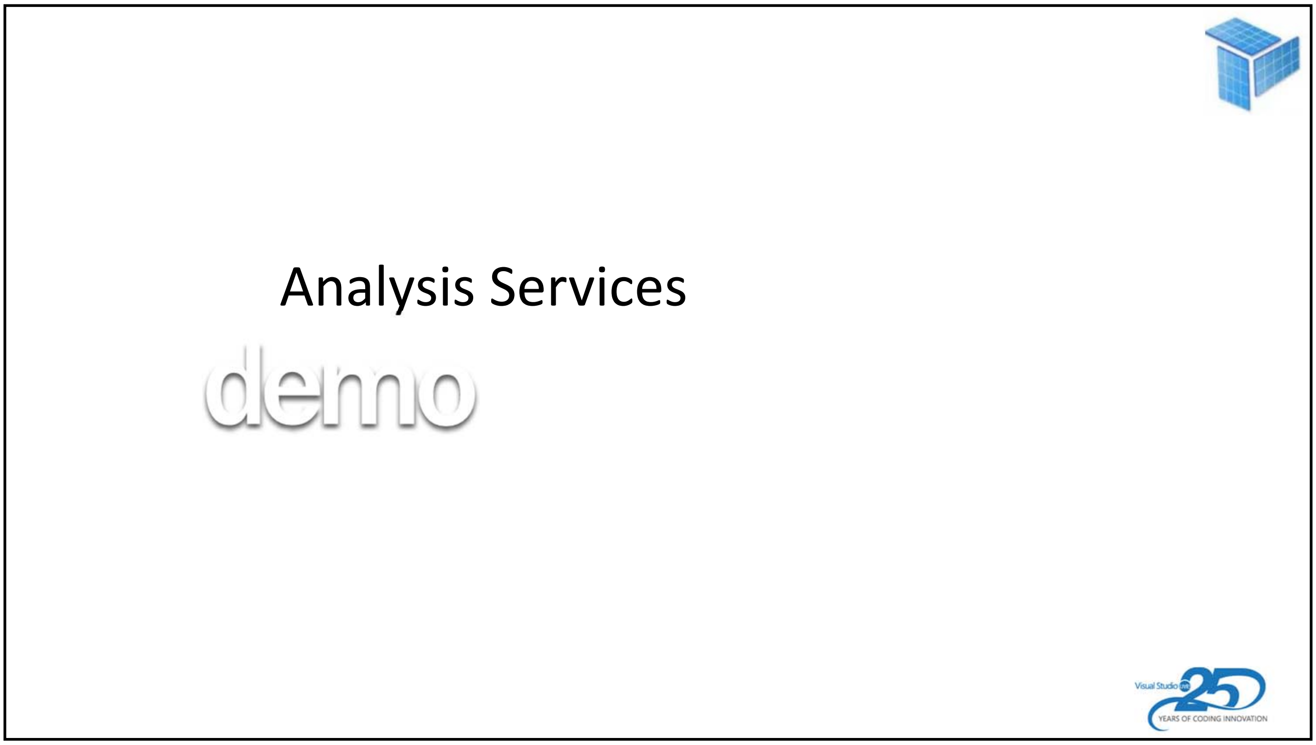
Analysis Services Tabular Mode

- SSAS Tabular Mode uses a columnar storage engine in place of a multidimensional one
- Must choose mode for SSAS instance at install time
- Can have default instance with one, named instance with the other
- Can create an SSAS Tabular database project by importing an Excel workbook with PowerPivot model
- SSAS tabular models support partitions, roles, translations, display folders



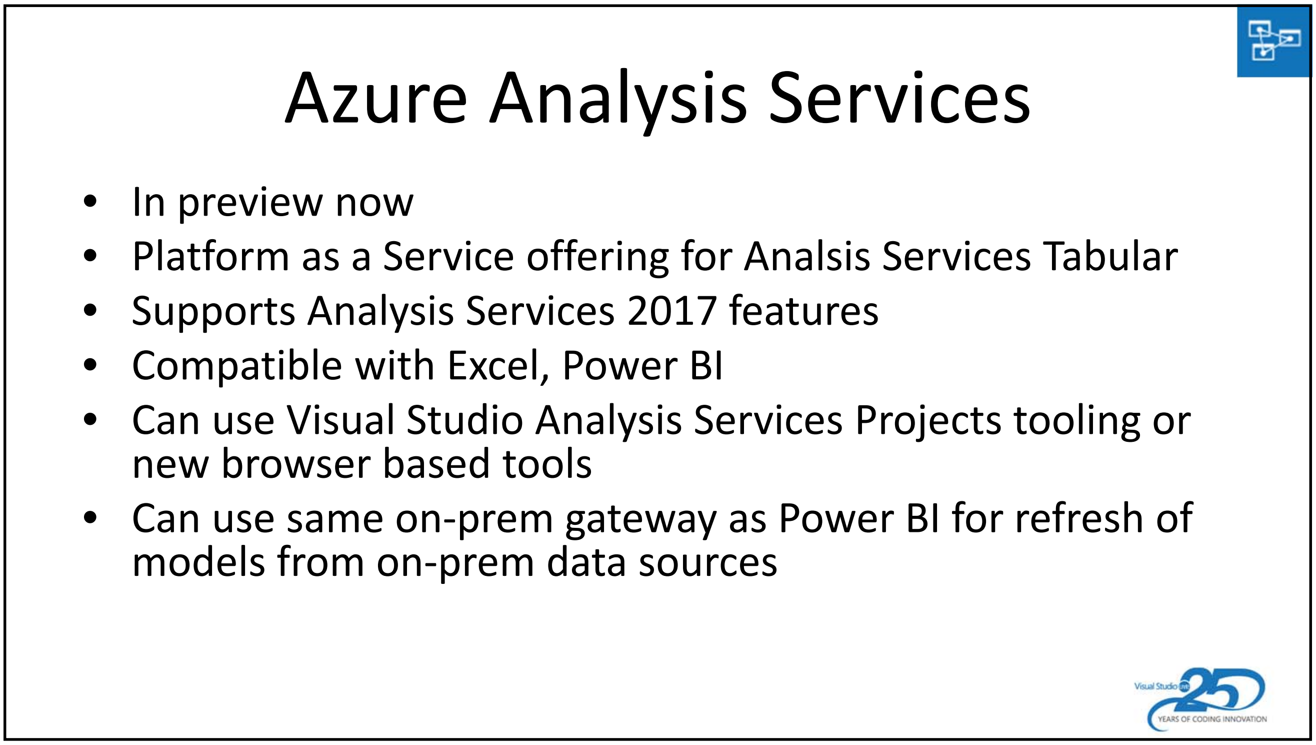
Calculated Columns and DAX

- Formula-based columns may be created
- Formula syntax is called DAX (Data Analysis eXpressions).
  - Not to be confused with MDX or DMX. Or DACs.
- DAX expressions are similar to Excel formulas
  - Work with tables and columns; similar to, but distinct from, worksheets and their columns (and rows)
- =FUNC('table name'[column name])
- =FUNCX('table name', <filter expression>)
- FILTER(Resellers,[ProductLine] = "Mountain")
- RELATED(Products[EnglishProductName])
- DAX expressions can be heavily nested



Analysis Services

demo



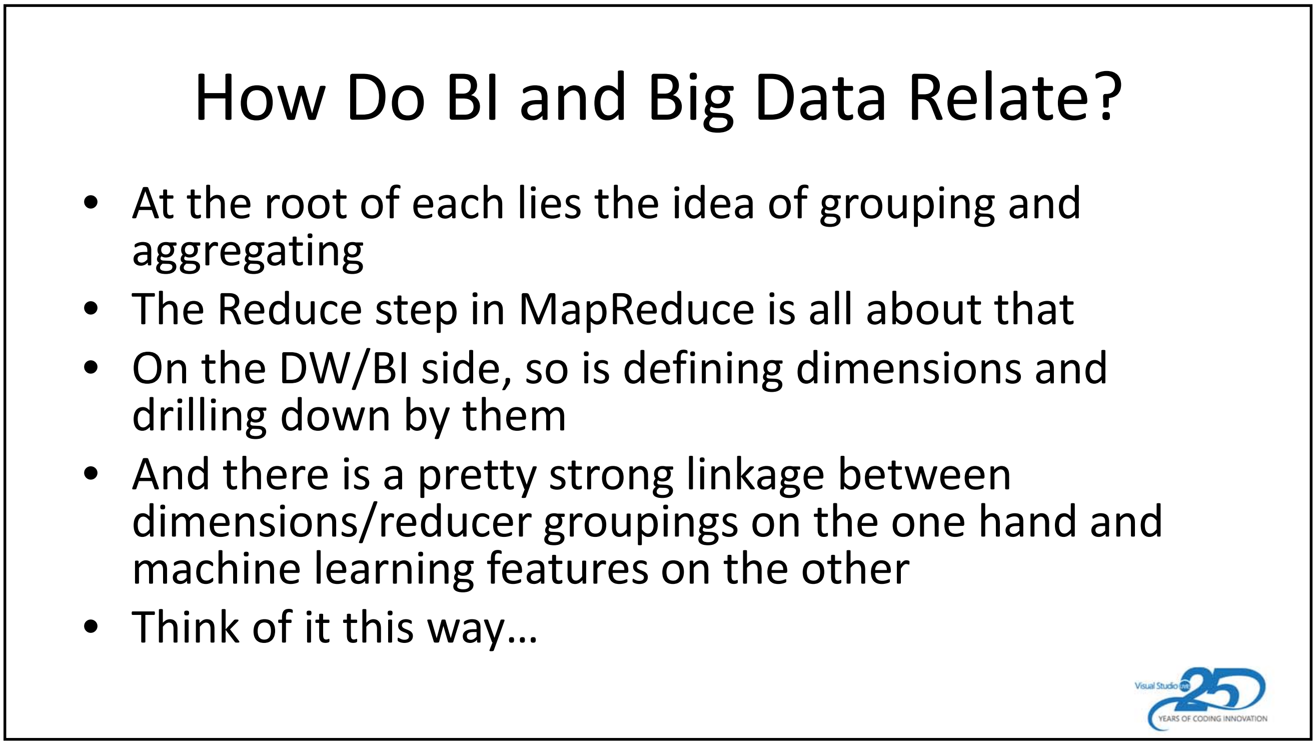
Azure Analysis Services

- In preview now
- Platform as a Service offering for Analsis Services Tabular
- Supports Analysis Services 2017 features
- Compatible with Excel, Power BI
- Can use Visual Studio Analysis Services Projects tooling or new browser based tools
- Can use same on-prem gateway as Power BI for refresh of models from on-prem data sources

Azure Analysis Services

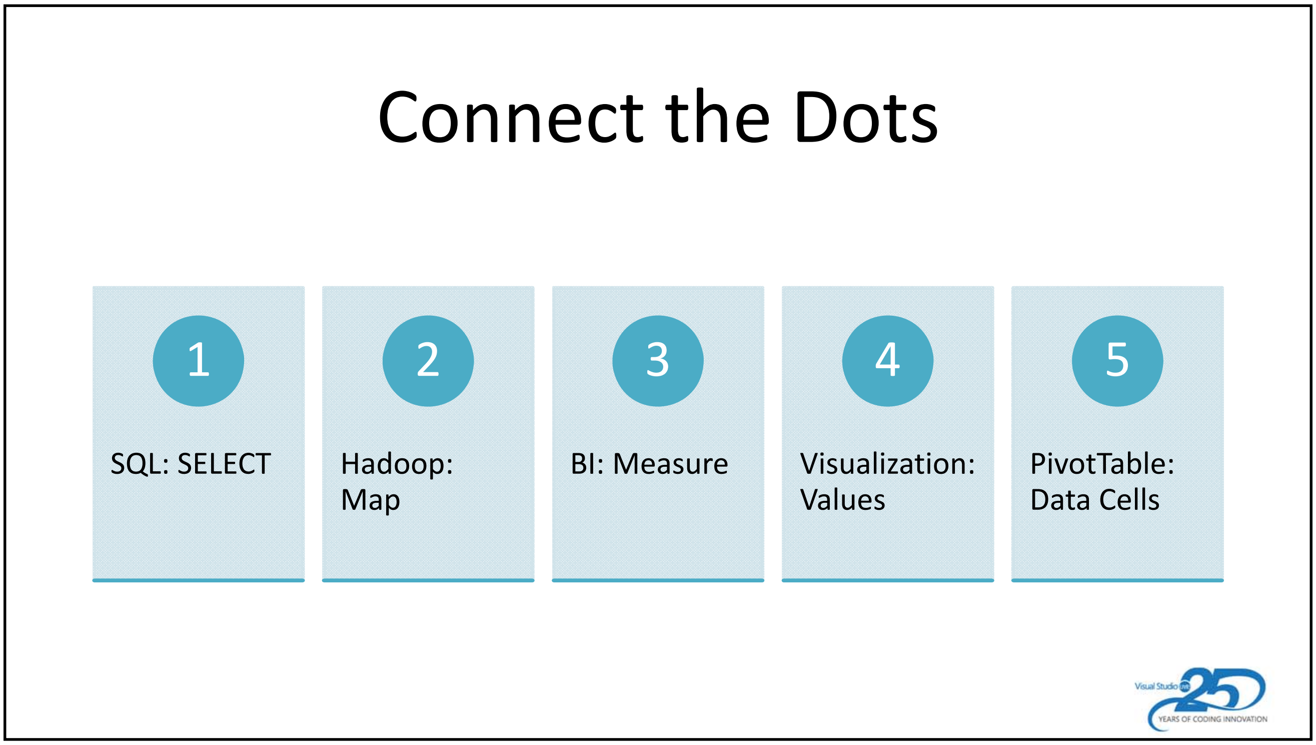
demo

CLOSING THOUGHTS



How Do BI and Big Data Relate?

- At the root of each lies the idea of grouping and aggregating
- The Reduce step in MapReduce is all about that
- On the DW/BI side, so is defining dimensions and drilling down by them
- And there is a pretty strong linkage between dimensions/reducer groupings on the one hand and machine learning features on the other
- Think of it this way…



Connect the Dots

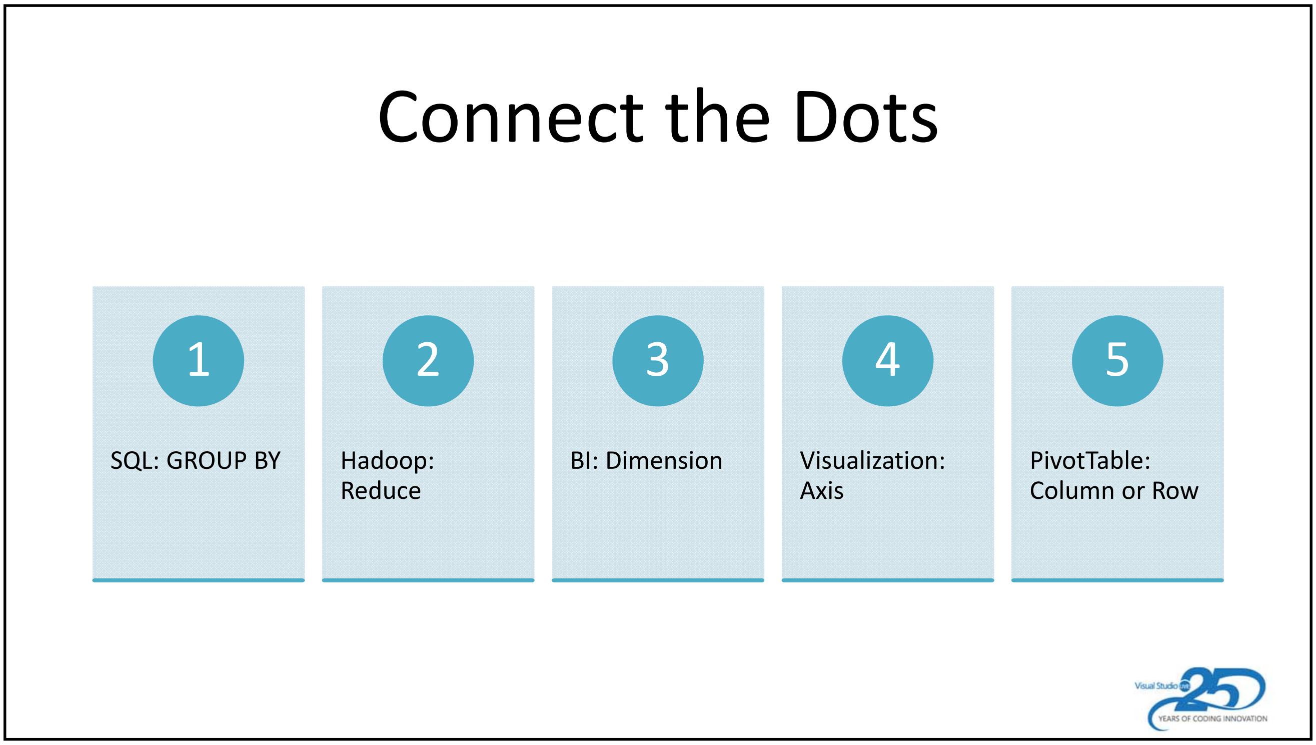
1 SQL: SELECT

2 Hadoop: Map

3 BI: Measure

4 Visualization: Values

5 PivotTable: Data Cells



Connect the Dots

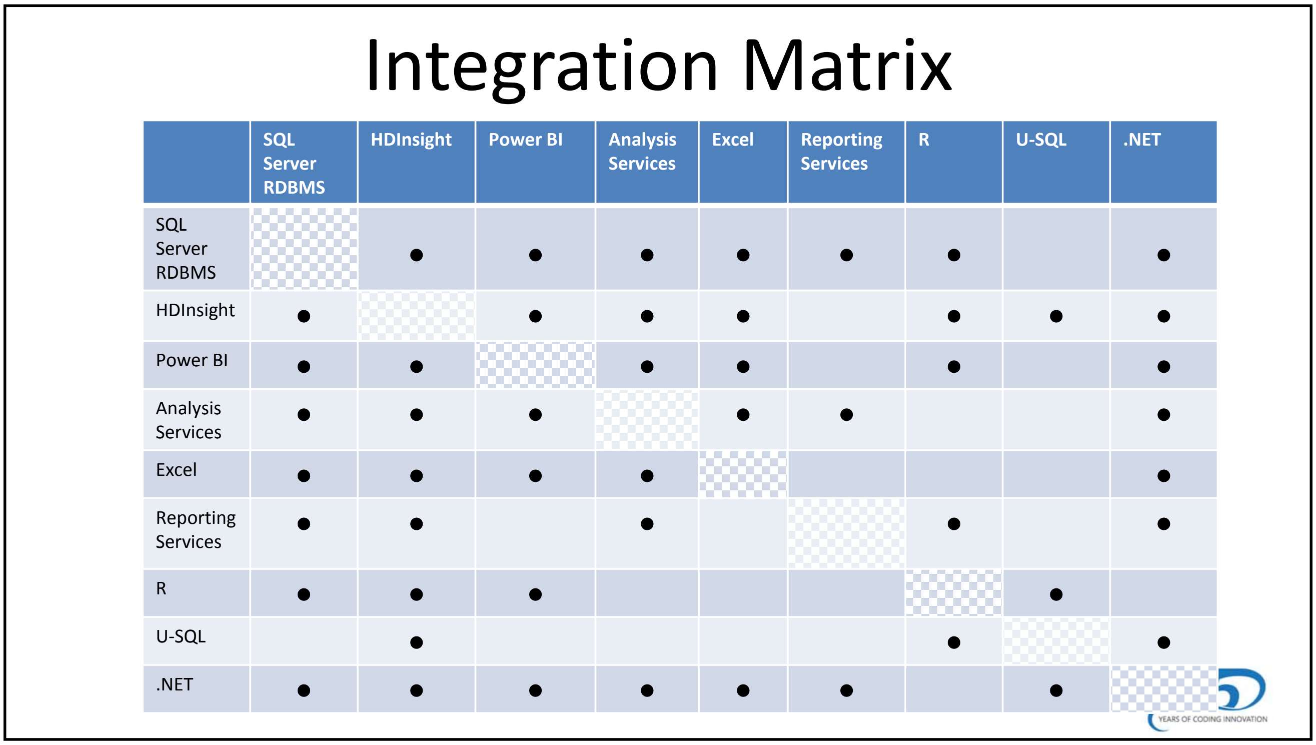
1 SQL: GROUP BY

2 Hadoop: Reduce

3 BI: Dimension

4 Visualization: Axis

5 PivotTable: Column or Row



Integration Matrix

	SQL Server RDBMS	HDInsight	Power BI	Analysis Services	Excel	Reporting Services	R	U-SQL	.NET
SQL Server RDBMS	▨	●	●	●	●	●	●		●
HDInsight	●	▨	●	●	●		●	●	●
Power BI	●	●	▨	●	●		●		●
Analysis Services	●	●	●	▨	●	●			●
Excel	●	●	●	●	▨				●
Reporting Services	●	●		●		▨	●		●
R	●	●	●				▨	●	
U-SQL		●					●	▨	●
.NET	●	●	●	●	●	●		●	▨