11.MAY ATHENS, GREECE

ATHENS AZURE BOOTCAMP 2019

One day full of **Microsoft Azure and the Cloud**

AzureHeads

autoexec.gr

SQLschool.gr

dotNETZone.gr
η δύναμη των ελλήνων προγραμματιστών

Microsoft Student Partners
CODE IS LIFE

student guru

Athens Global Azure Bootcamp 2019

Big Data analytics:
Finding diamonds in the rough with Azure

Christos Charmatzis
@T.A. Geoforce

DATA TEAM

Agenda

- Introduction
- When we have a Big Data problem
- Finding the best solution for our Big Data
- Working inside the Data Team
- Extract the true value of our data

Introduction

What is Big Data?
"Big Data" is a field that treats ways to analyze, systematically extract information from, or otherwise deal with data sets that are too large or complex to be dealt with by traditional data-processing application software.

Source: Wikipedia

The concept gained momentum in the early 2000s when industry analyst Doug Laney articulated the now-mainstream definition of big data as the three Vs:

1. Volume
2. Velocity
3. Variety

And because everything is relative

What is today’s small (1TB), was yesterday’s big.....
And what is today’s big(100TB) is tomorrow’s small….

(*we use 100TB, because is the dataset size of Sort Benchmark competition* *http://sortbenchmark.org/*)

When we have a Big Data problem?

Example 1

- 450GB Datasets
- Machine (M32ls Instance , 32 VCPU, 256 GiB RAM, 1,024 GiB Storage, ~€2,122.3736/month)
- Enterprise Database (e.g. SQL Server)
- Aggregation, Statistics, Summaries

STAY WHERE YOU ARE

Example 2

- 3TB Datasets
- Machine (M32ls Instance , 32 VCPU, 256 GiB RAM, 1,024 GiB Storage, ~€2,122.3736/month)
- Enterprise Database (e.g. SQL Server)
- Aggregation, Statistics, Summaries

UPGRADE STORAGE

Example 3

- 10TB Dataset
- Aggregation, Statistics, Summaries, Transformations etc

GO TO THE CLOUD

Example 4

- 450GB Dataset
- Machine (M32ls Instance , 32 VCPU, 256 GiB RAM, 1,024 GiB Storage, ~€2,122.3736/month)
- Enterprise Database (e.g. SQL Server)
- Transformations

GO TO THE CLOUD

Big Data Infrastructure comparison

DB in premise

- Initial release < 2014
- Supported programming languages: almost every programming language
- Performance keys: Indexes

!==

Spark Cluster

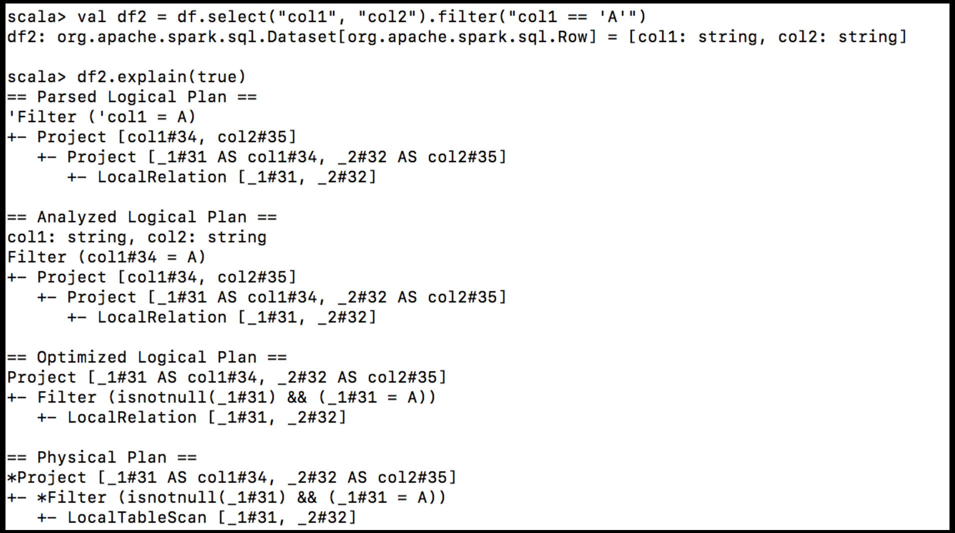
- Initial release > 2014
- Supported programming languages: Java, Scala, Python, R, Julia
- Performance keys: Partitioning

Big Data Performance

In Spark always:

- use “df.explain(true)”
- Or check the DAG!

Every time a block is changing the data are repartitioning!!!



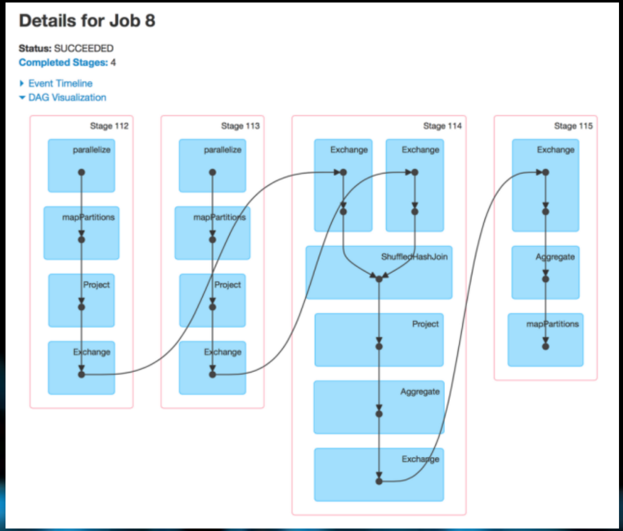
```
scala> val df2 = df.select("col1", "col2").filter("col1 == 'A'")
df2: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [col1: string, col2: string]

scala> df2.explain(true)
== Parsed Logical Plan ==
'Filter ('col1 = A)
+- Project [col1#34, col2#35]
   +- Project [_1#31 AS col1#34, _2#32 AS col2#35]
      +- LocalRelation [_1#31, _2#32]

== Analyzed Logical Plan ==
col1: string, col2: string
Filter (col1#34 = A)
+- Project [col1#34, col2#35]
   +- Project [_1#31 AS col1#34, _2#32 AS col2#35]
      +- LocalRelation [_1#31, _2#32]

== Optimized Logical Plan ==
Project [_1#31 AS col1#34, _2#32 AS col2#35]
+- Filter (isnotnull(_1#31) && (_1#31 = A))
   +- LocalRelation [_1#31, _2#32]

== Physical Plan ==
*Project [_1#31 AS col1#34, _2#32 AS col2#35]
+- *Filter (isnotnull(_1#31) && (_1#31 = A))
   +- LocalTableScan [_1#31, _2#32]
```



Details for Job 8

Status: SUCCEEDED

Completed Stages: 4

Finding the best solution for our Big Data problem

- Hadoop on a cluster of Azure Virtual Machines
- Azure HDInsights (Clusters as-a-service)
- Azure Databricks
- Azure Data Factory (New & Improved!!!!)
- Azure Data Lake Analytics (Queries as-a-service)

Big Data in Azure: Storage

Azure Blob Storage

- Object Storage
- General purpose (files & workloads)

Azure Data Lake

- Hierarchical file system
- Optimized for analytics workloads

Azure Data Lake (Gen.2)

- Multi-modal storage
- Optimized for analytics workloads

Big Data in Azure: Storage

Azure Blob Storage

wasb[s]://containername@accountname.blob.core.windows.net/file.csv

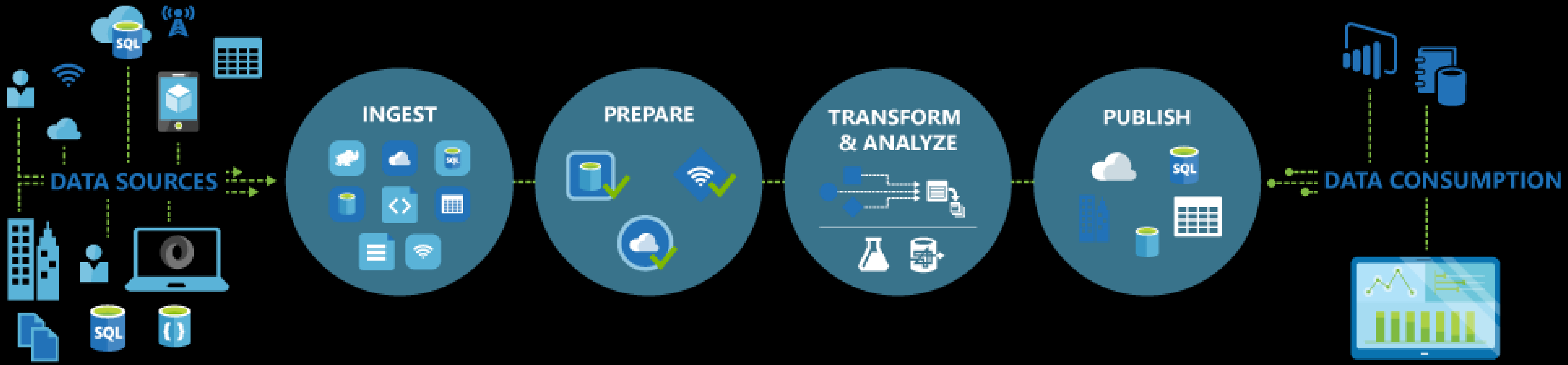
Azure Data Lake

abfs[s]://filesystemname@accountname.dfs.core.windows.net/file.csv

Azure Data Lake (Gen.2)

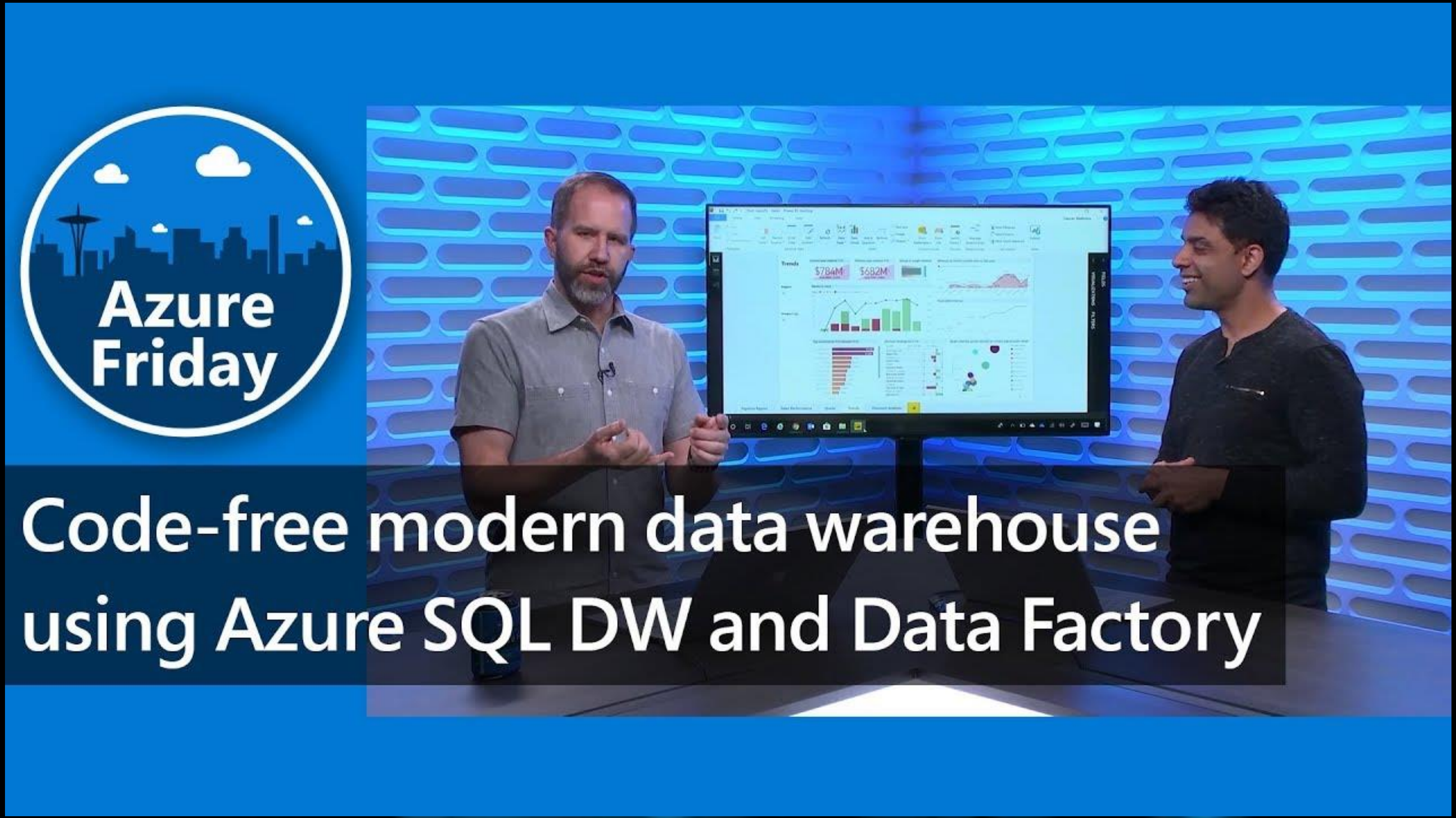
- Endpoint: object store access Blob API using wasb[s]://
- Endpoint: file system access ADLS Gen 2 API using abfs[s]://

Azure Data Factory



A managed could service for building & operating data pipelines.

Azure Data Factory


Azure Friday
Code-free modern data warehouse using Azure SQL DW and Data Factory
$784M
$682M

Source: https://channel9.msdn.com

Azure Data Factory (ADF)

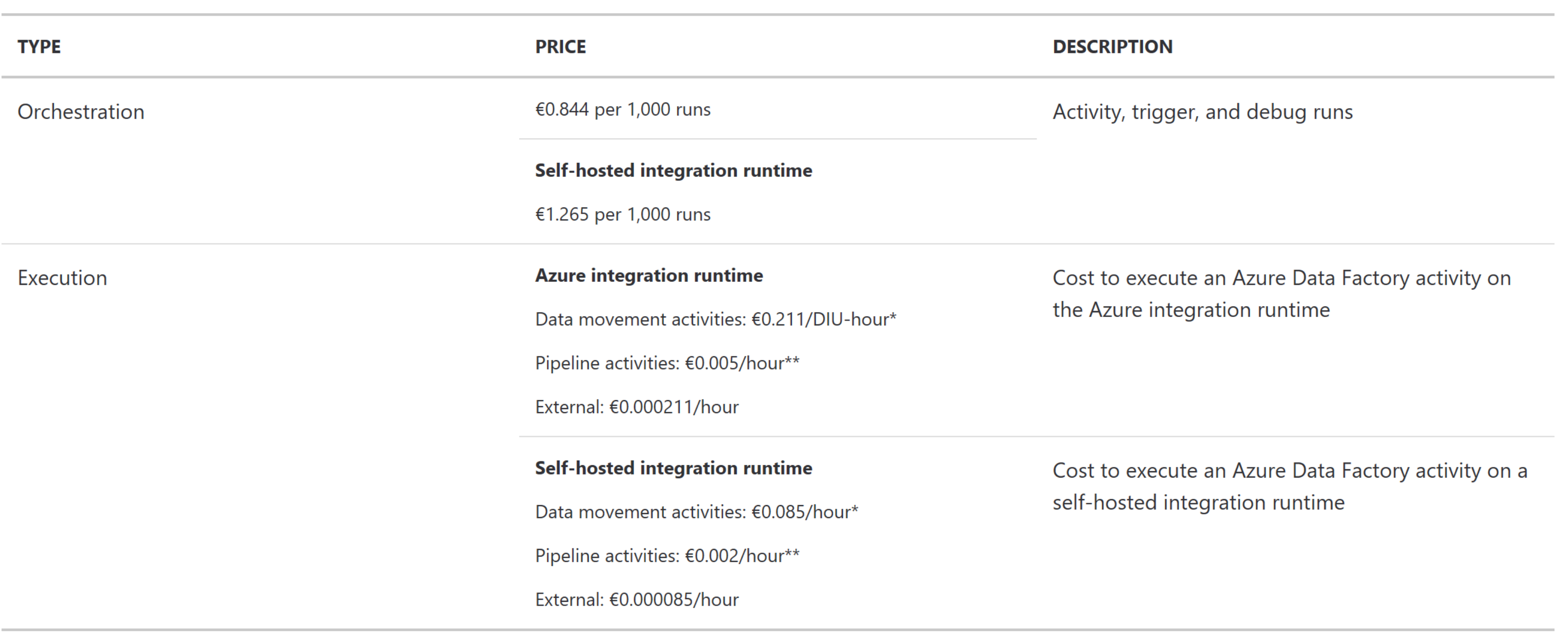
DEMO

Why do we need tools like ADF?

85% of the working time is on data wrangling!!!

ADF Pricing

https://azure.microsoft.com/en-us/pricing/details/data-factory/



TYPE	PRICE	DESCRIPTION
Orchestration	€0.844 per 1,000 runs	Activity, trigger, and debug runs
	Self-hosted integration runtime €1.265 per 1,000 runs	
Execution	**Azure integration runtime** Data movement activities: €0.211/DIU-hour* Pipeline activities: €0.005/hour** External: €0.000211/hour	Cost to execute an Azure Data Factory activity on the Azure integration runtime
	Self-hosted integration runtime Data movement activities: €0.085/hour* Pipeline activities: €0.002/hour** External: €0.000085/hour	Cost to execute an Azure Data Factory activity on a self-hosted integration runtime

Tip: Look out!!! The data reads and writes are the most expensive in Big Data Analytics

Working inside the Data Team

Somewhere, inside a meeting room….
I will run the whole thing, again
We must compare the results with last year's...
Don't we have somewhere that report?
Yes, there're in a folder, inside a VM, inside John's PC...
No, we have uploaded them in blob storage... I don't remember

Rembrandt (1662). The Sampling Officials (Dutch: De Staalmeesters)

Metadata area

With Data comes problems….

With Big Data comes Bigger Problems!!!

Like….

- Many datasets
- Frequently updates
- Many fields
- Many users

Where do I keep the metadata?

- Azure Data Catalogue
- DataBricks Delta Lake
- Create your own meta-portal

Be aware, always use metadata standards (ISO, Dublin Core, MPEG-7 …)

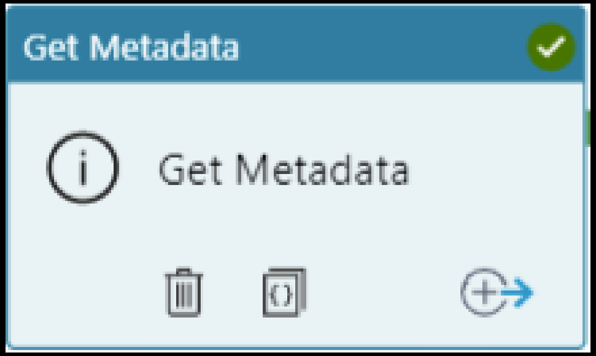
More info: https://en.wikipedia.org/wiki/Metadata_standard#Available_metadata_standards

Azure Data Factory Metadata

This activity allows for collecting metadata about Azure Data Factory.

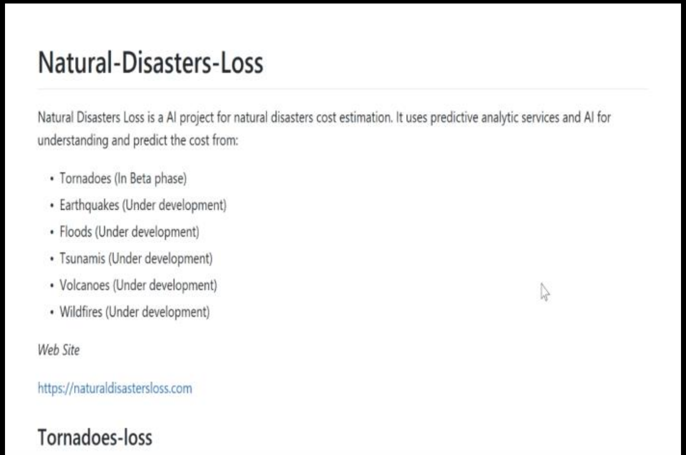
Get Metadata activity supports:

- itemName
- itemType
- Size
- Created
- lastModified
- childItems
- contentMD5
- Structure
- columnCount
- exists



Extract real value from the data

Visualize data | Write good experiments | Share results



Natural-Disasters-Loss

Natural Disasters Loss is a AI project for natural disasters cost estimation. It uses predictive analytic services and AI for understanding and predict the cost from:

- Tornadoes (In Beta phase)
- Earthquakes (Under development)
- Floods (Under development)
- Tsunamis (Under development)
- Volcanoes (Under development)
- Wildfires (Under development)

Web Site

https://naturaldisastersloss.com

Tornadoes-loss

Share

And we just scratched the surface of that…

Streaming
Hadoop
Databases
AI
Excel
Scala
BigData
R
U-SQL
Kafka
Python
Java
Jupiter
Zepellin
Data
PowerBI
Machine Learning

Conclusions

- For ETL projects from in premise to cloud use Azure Data Factory
- The size isn’t always the problem in your case
- Velocity isn’t only on the code side, you HAVE to know your data
- Create METADATA

Thank U

Qs+As

Please evaluate:
http://bit.ly/AAB2019Evaluation