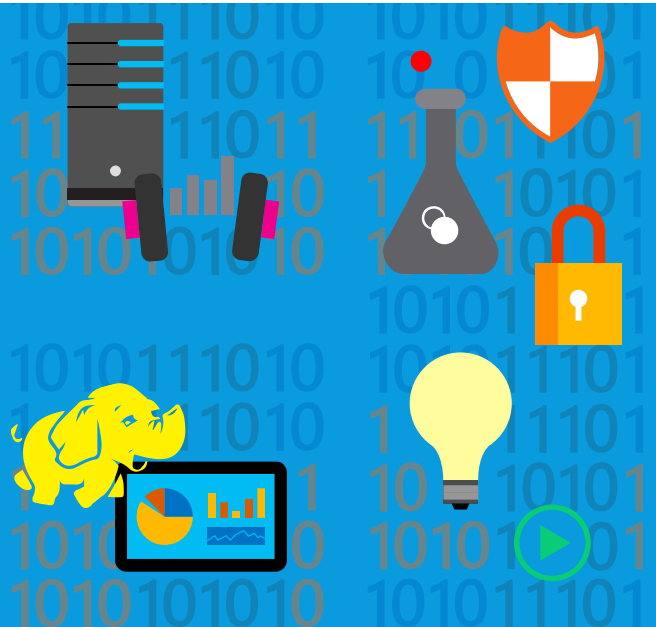
AZURE HDINSIGHT

Azure Machine Learning Track

Marek Chmel

2017 Global Azure BOOTCAMP

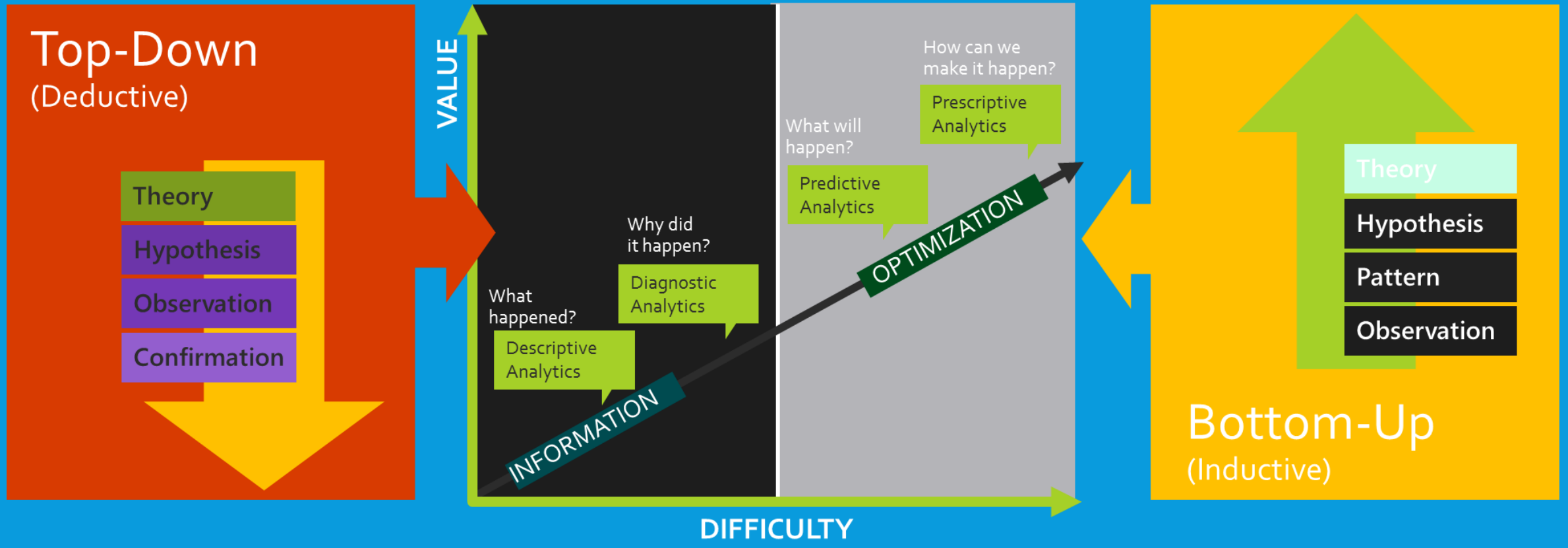


SESSION AGENDA

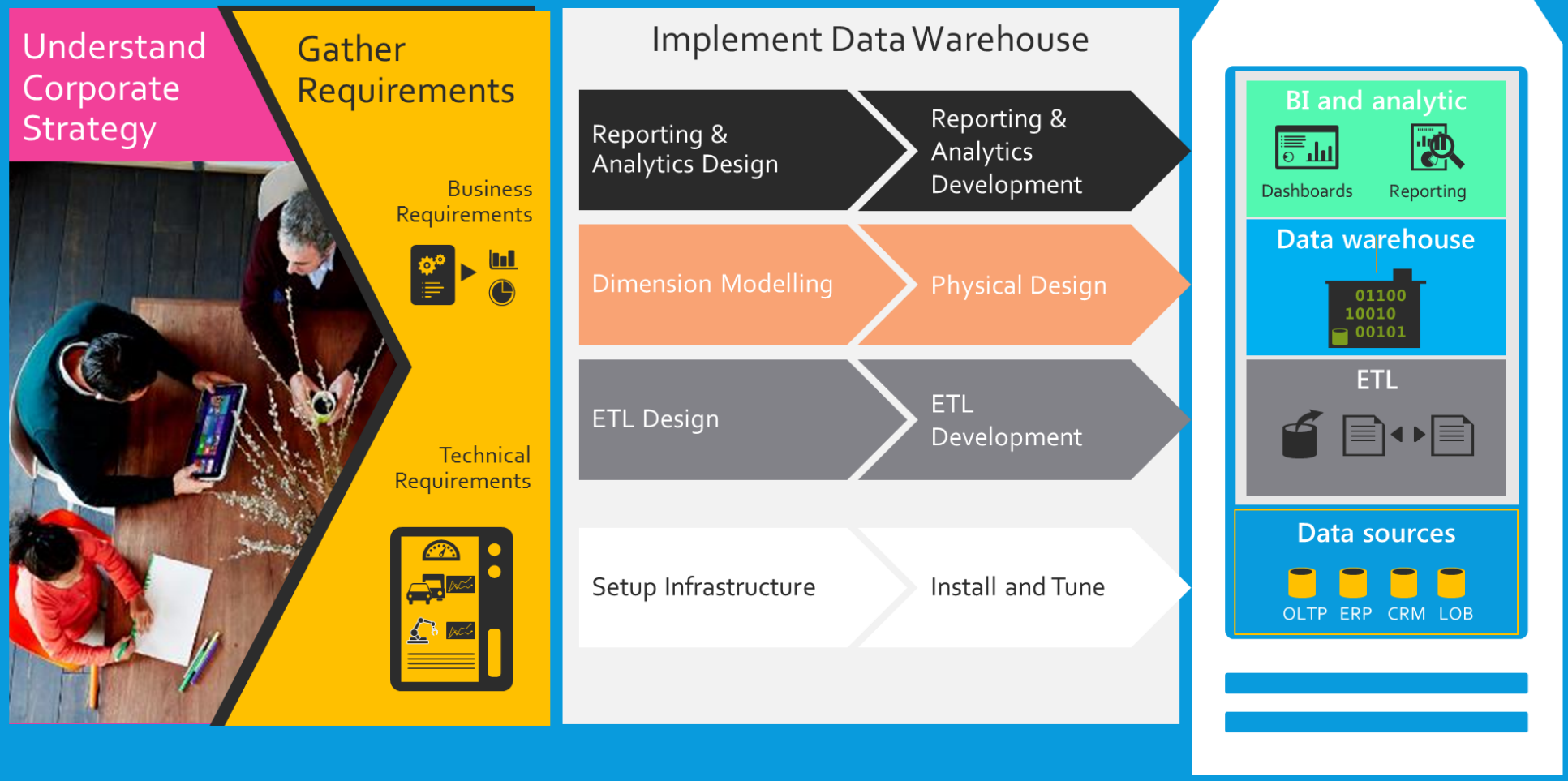
- Understanding different scenarios of Hadoop
- Building an end to end pipeline using HDInsight
- Using in-memory techniques to analyze data interactively

BIG DATA VS. TRADITIONAL DW

TWO APPROACHES TO INFORMATION MANAGEMENT FOR ANALYTICS: TOP-DOWN + BOTTOM-UP



DATA WAREHOUSING USES A TOP-DOWN APPROACH



THE "DATA LAKE" USES A BOTTOM-UP APPROACH

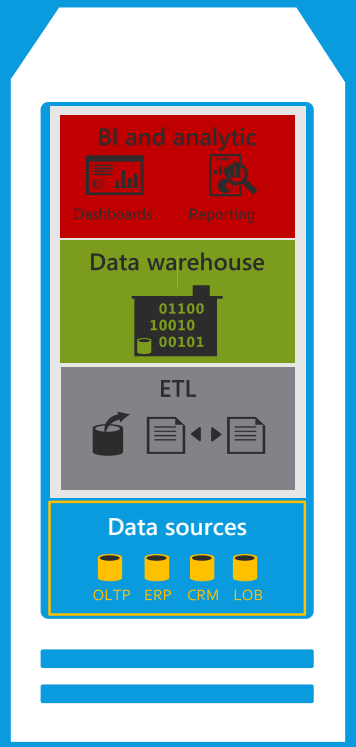
DATA LAKE + DATA WAREHOUSE BETTER TOGETHER

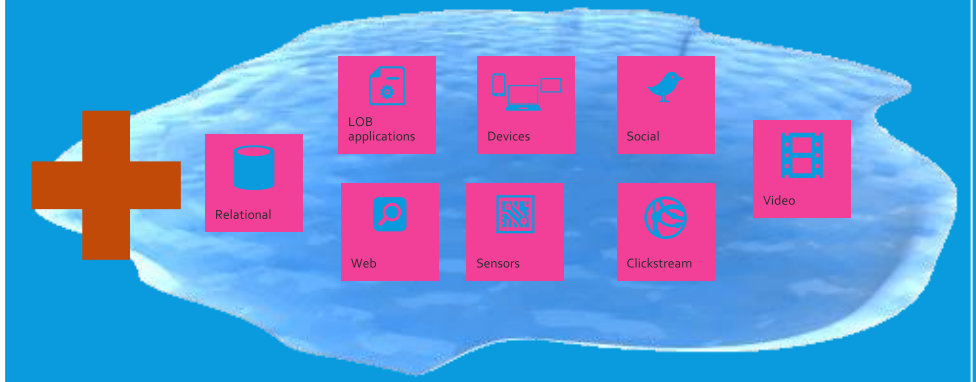
What happened?

What is happening?

Why did it happen?

What are key relationships?





What will happen?

What if?

How risky is it?

What should happen?

What is the best option?

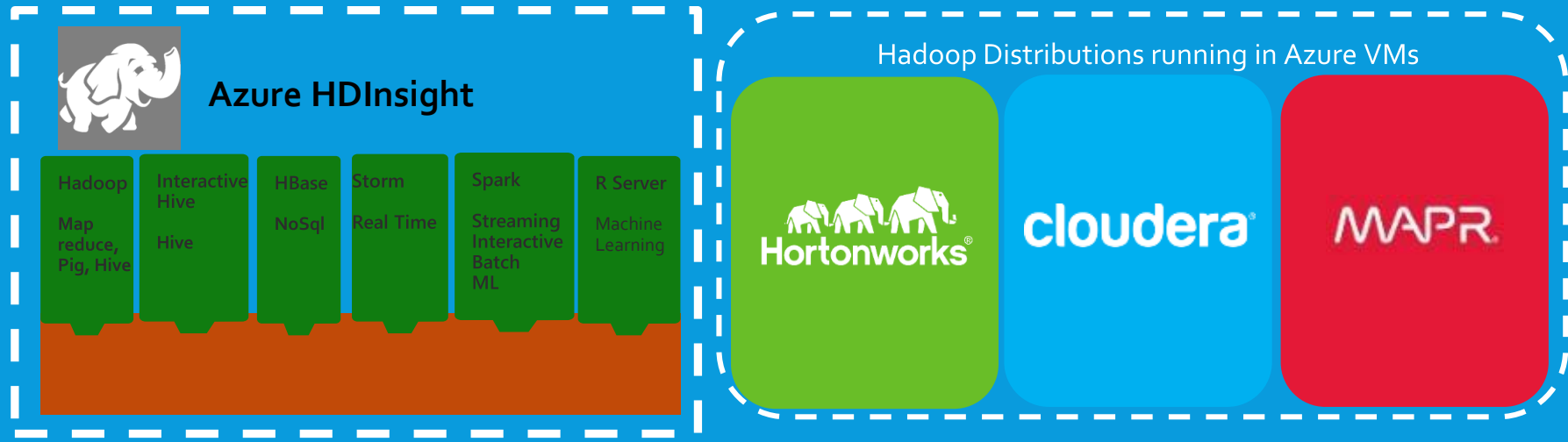
How can I optimize?

WHAT IS HDINSIGHT

MICROSOFT HADOOP STACK

Analytics

Storage



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

Hadoop clusters have grown by 60% in the last 2 years

89% of enterprise users consider Hadoop as opportunity for innovation

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

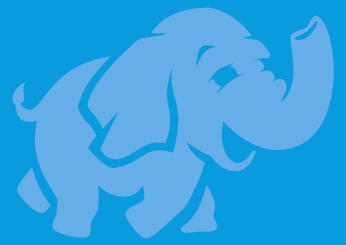
Hadoop is shifting from a buzzword to a real production service

Ownership is shifting from department teams to Central IT.

AZURE HDINSIGHT

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

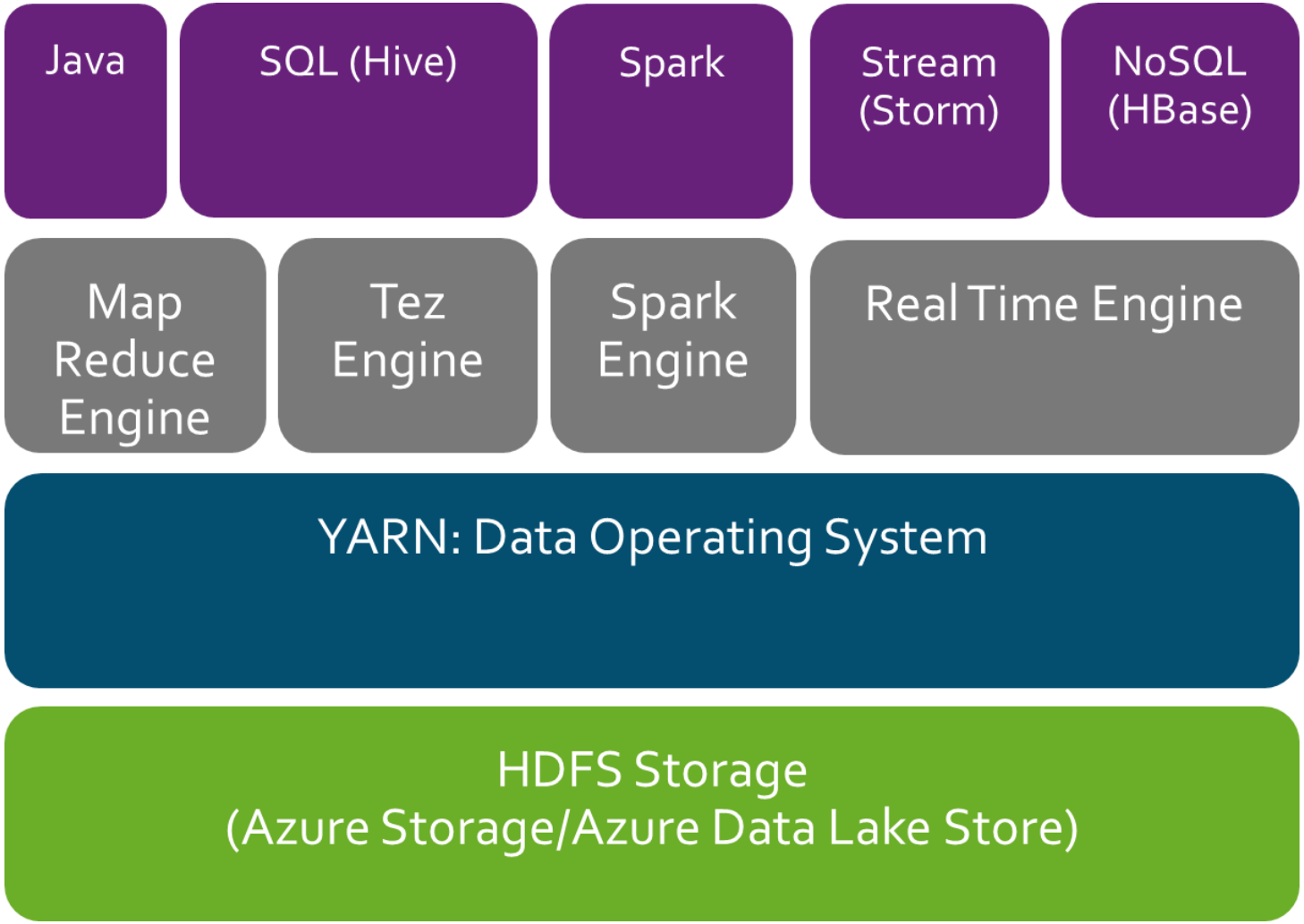
Hadoop and Spark
as a Service on Azure



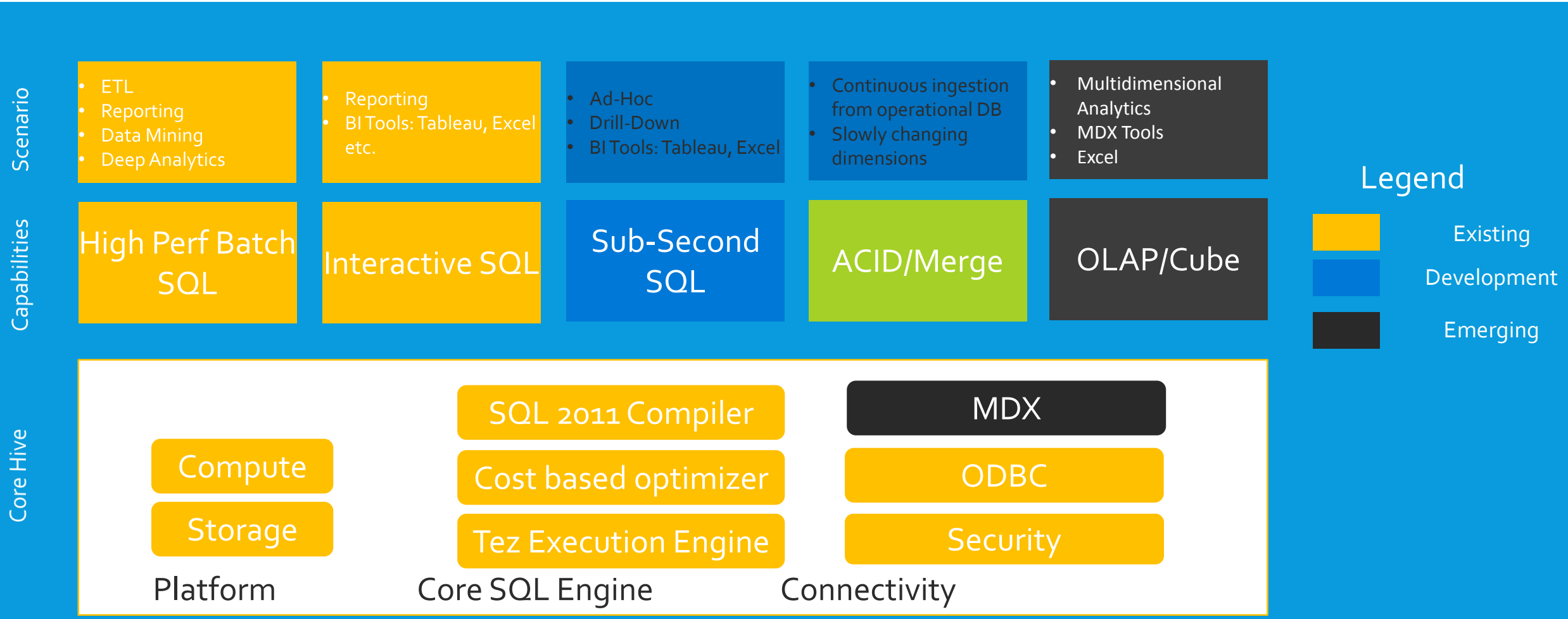
HDINSIGHT WORKLOADS

- Hadoop
  - Batch: Hive and MapReduce
  - Interactive Hive using LLAP (New – just launched)
- HBase (NoSQL)
- Storm (Streaming)
- Spark (Interactive)

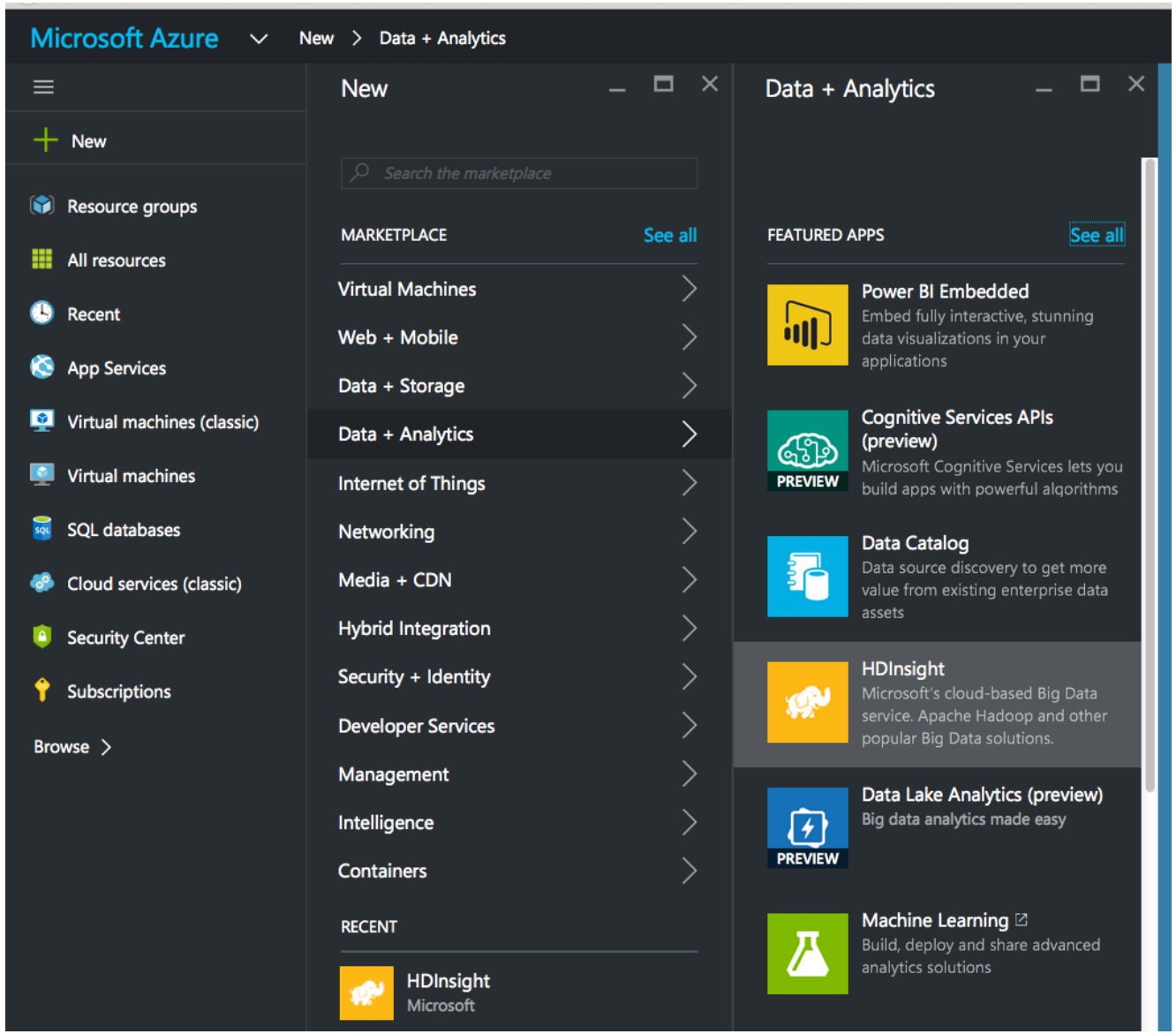
HDINSIGHT ARCHITECTURE



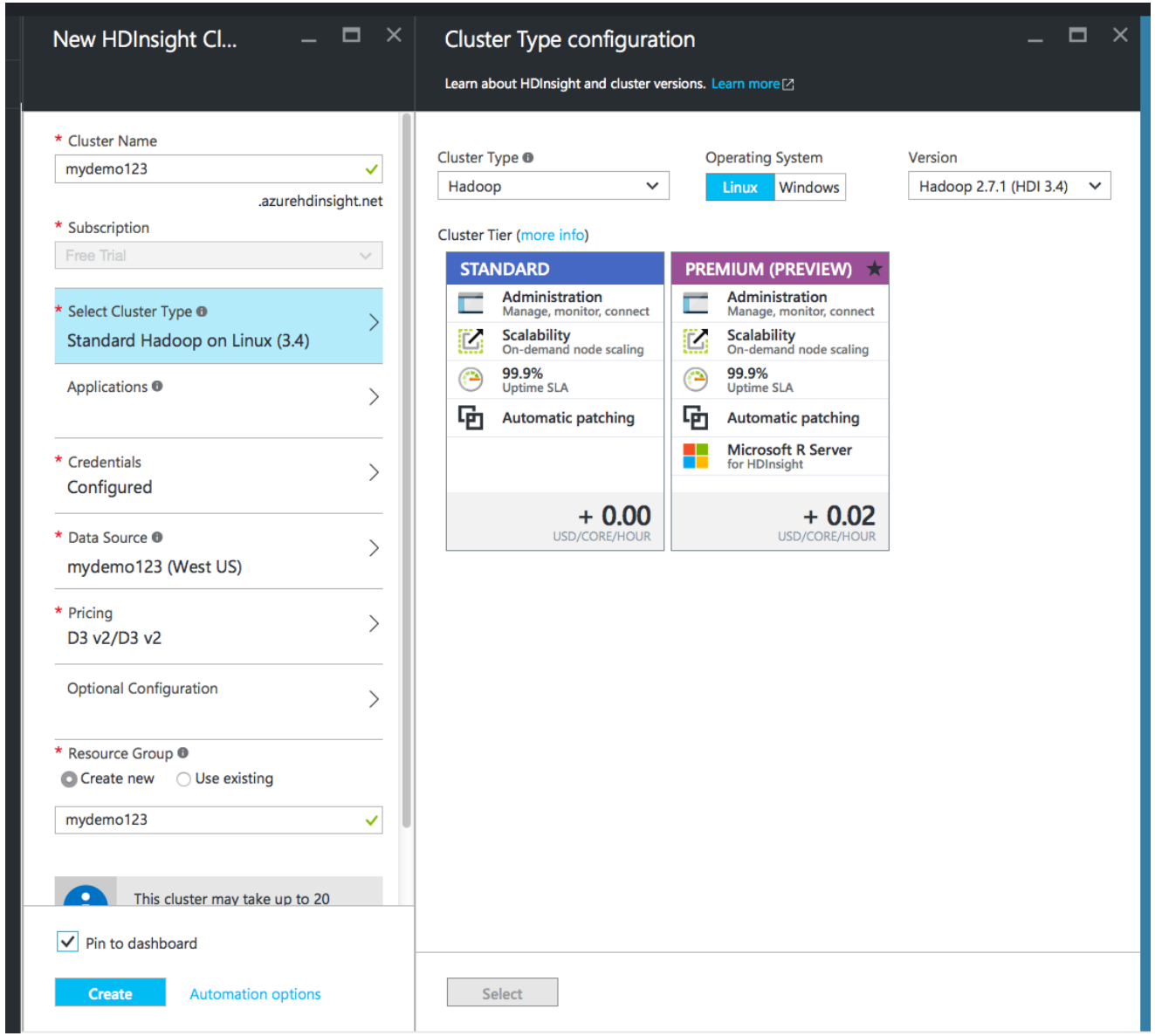
INTRO TO HIVE



CREATING HDINSIGHT CLUSTER

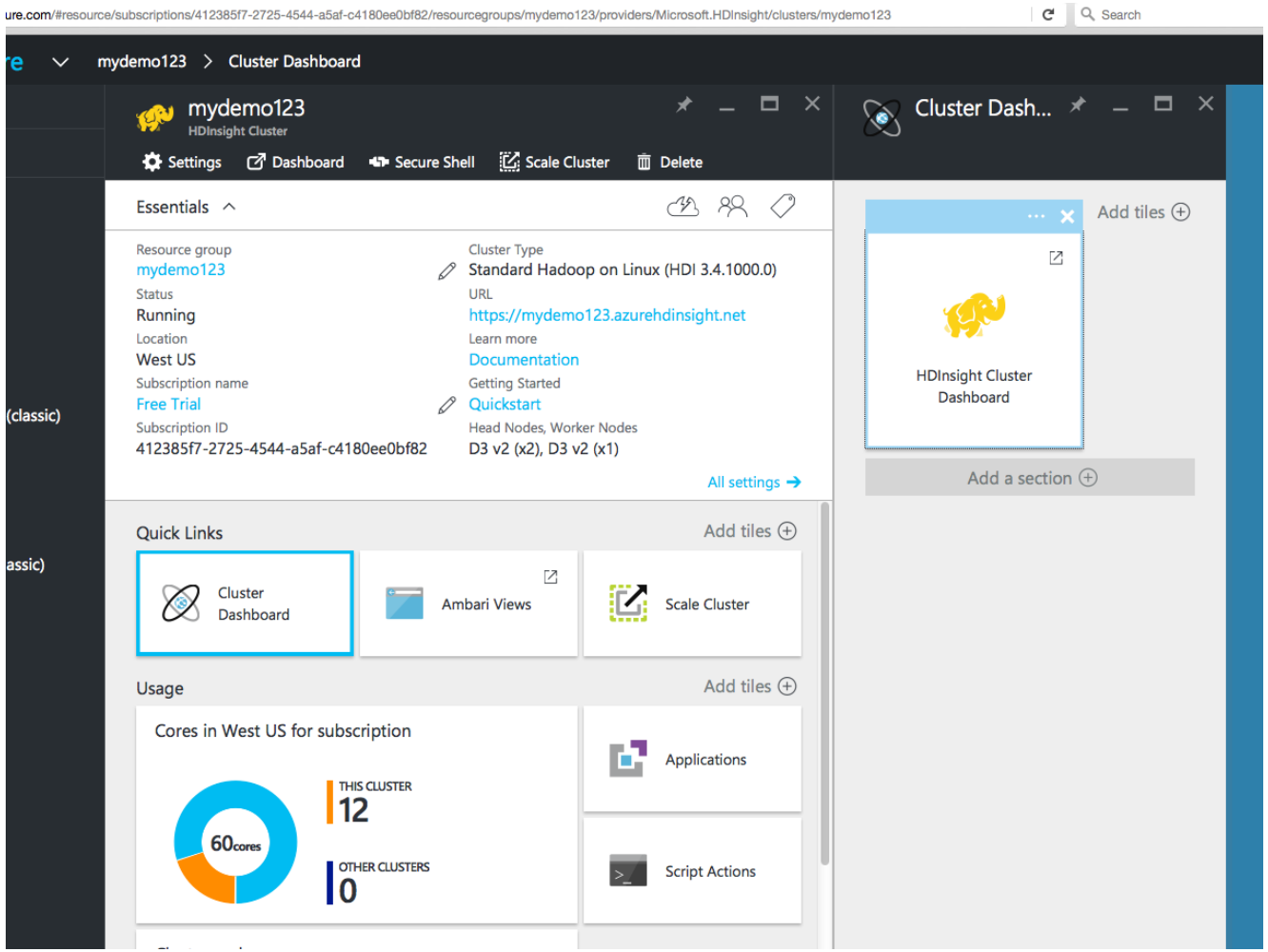


CREATING HDINSIGHT CLUSTER



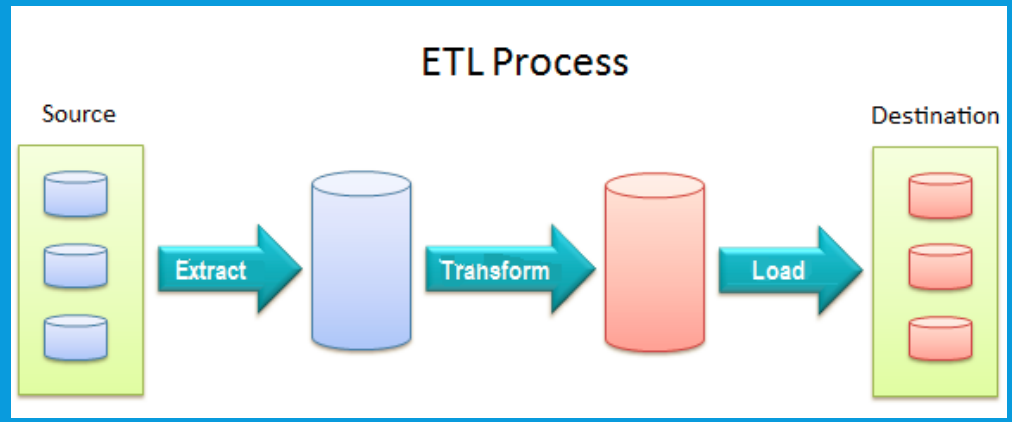
CLUSTER DASHBOARD

- Ambari dashboard which you can use with Hortonworks HDP



TYPICAL HADOOP SCENARIOS

- ETL
  - Data ingested from various sources
  - Transformed and cooked to structured data
  - Then loaded into a DB for querying
  - Typically batch scenario
- BI Scenarios
  - Used by business analyst for ad-hoc querying
  - Requires interactive response



BUILDING AN ENTERPRISE DW USING HADOOP

- Planning
  - Cluster planning
  - Cluster Deployment model
- Development
  - Author and Debug Queries
  - Optimize queries
- Deployment
  - Use ADF/Oozie to schedule and productionalize your jobs
  - Monitor and manager cluster using Ambari
- Connecting with BI tools
  - Create tables on ORC data from shared storage account
  - Have BI tools connect to cluster using ODBC driver

CLUSTER PLANNING

- Understand requirements
  - What is scenario?
  - What is SLA?
  - What is budget?
  - How often?
  - Who is the customer?
- Type of cluster
  - Production, Dev or Test?
  - On-demand vs. persistent?
  - Custom vs. default metastore?
  - Security model?
- Trade-offs
  - Single or multi tenant?
  - CPU or Memory bound?

CLOUD DEPLOYMENT MODELS

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

Why use Cluster as a Service?

- Pay only for time the cluster was actually used
- Since data & metadata is persisted, experience is as if the cluster was never deleted

QUERY AUTHORING

Ambari Views

- Provides graphical UX for authoring and debugging Hive queries
- Pros: One of the few tools that can be used to debug Tez queries

Visual Studio

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

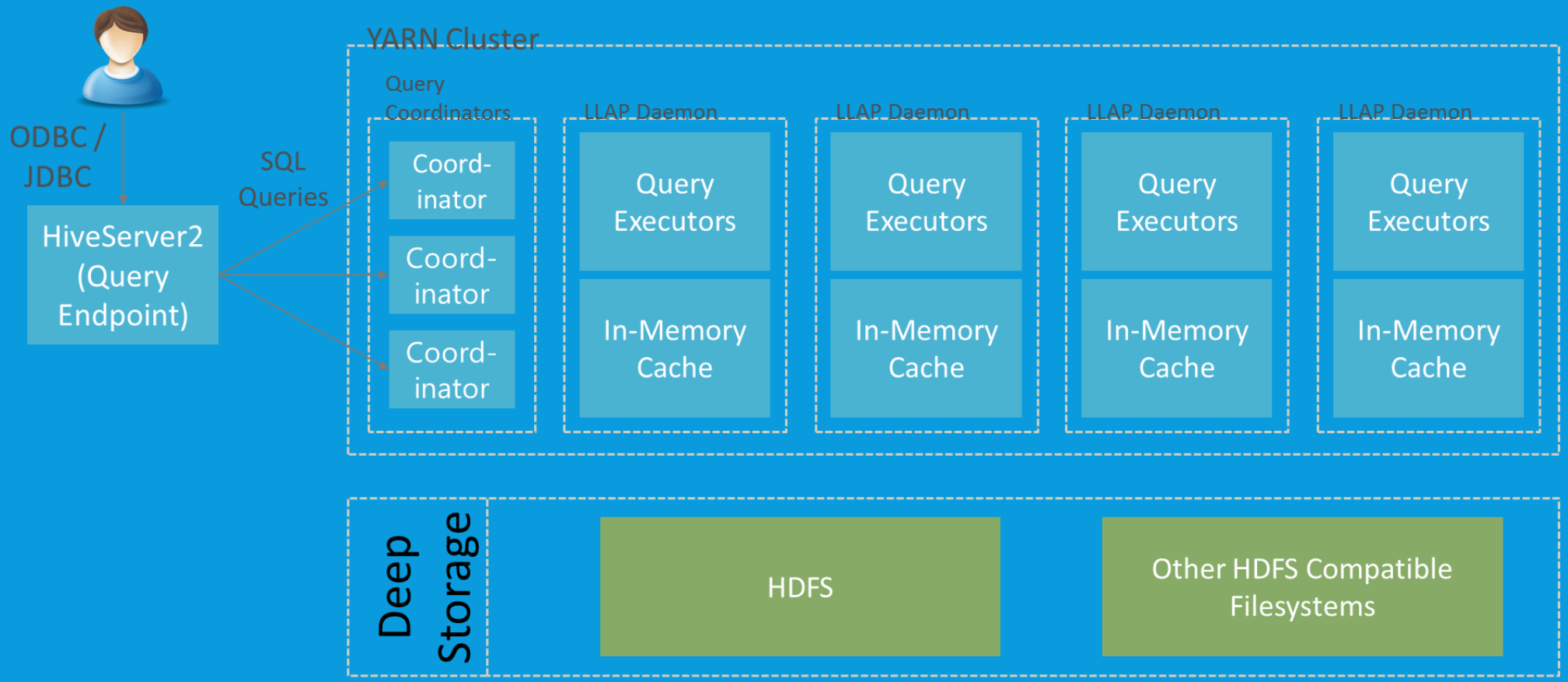
Command-Line

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

Beeline

- Command line shell that works with HiveServer2.
- Pros: Very thin JDBC client

INTRODUCING HIVE LLAP: MAKING HIVE INTERACTIVE



HIVE LLAP

- Interactive Querying through in-memory compute
- 10x-25x faster than using Hive1
- Allows multiple users to run queries simultaneously
- Provides enterprise class security
- Separate capacity for ETL and EDW scenarios
- Integration with world class BI tools

SCALING FOR BIG DATA WORKLOADS

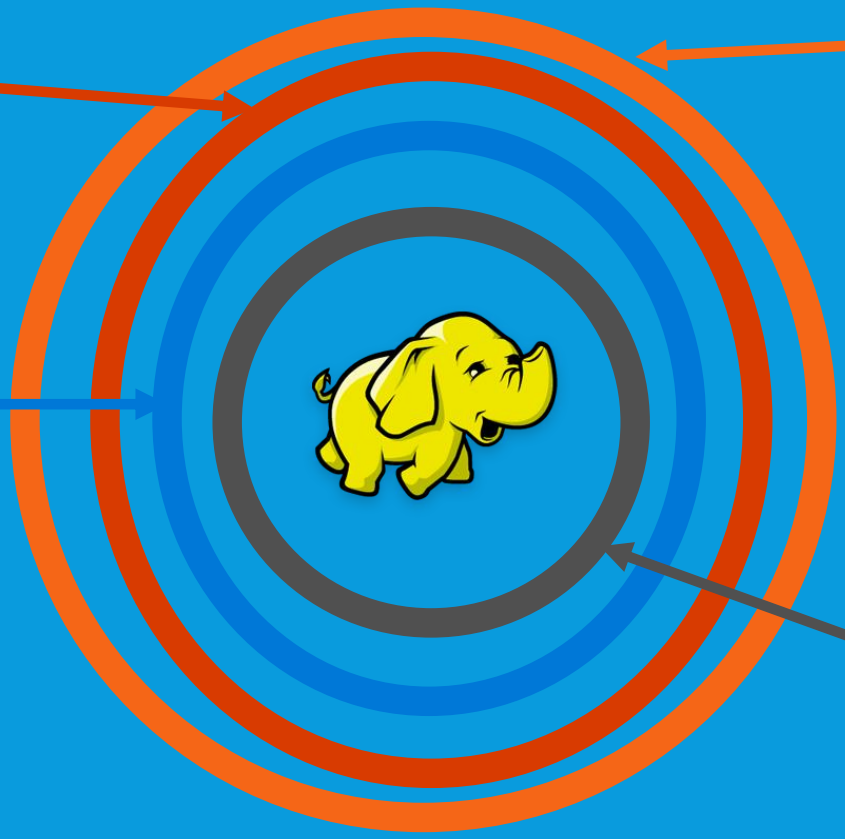
- Challenges
  - Improving High Availability
  - Elastic scaling
  - Ability to scale to multiple users

- How HDInsight helps with scaling
  - Platform Availability improvements
  - Ability to scale during and after cluster creation
  - Ability to create Edge Nodes

HDINSIGHT SECURITY – RINGS OF DEFENSE

Authentication
Kerberos
Active Directory

Authorization
Hive policies
HBase policies
File and Folder level ACLS

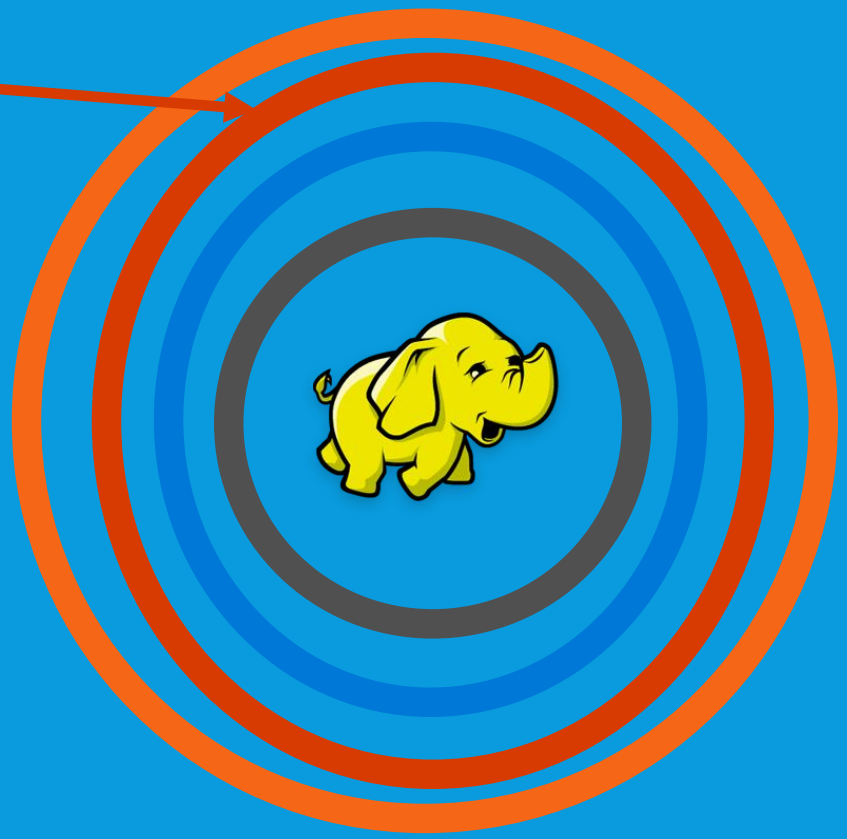


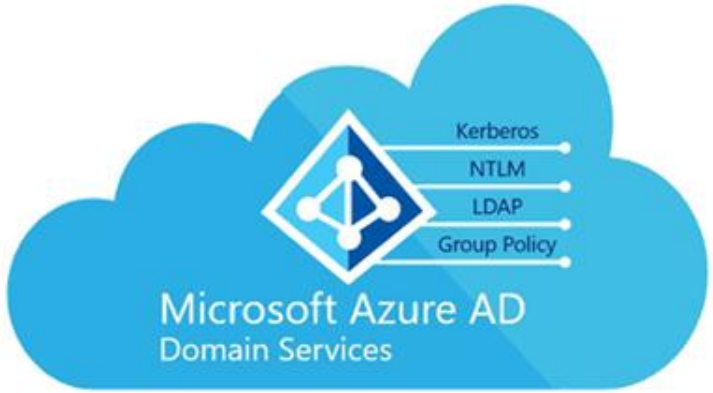
Perimeter Level Security
Virtual Network
Network Security (i.e. Firewalls)
Gateway

Data Security
Encryption @ Rest

INTEGRATION WITH AZURE ACTIVE DIRECTORY

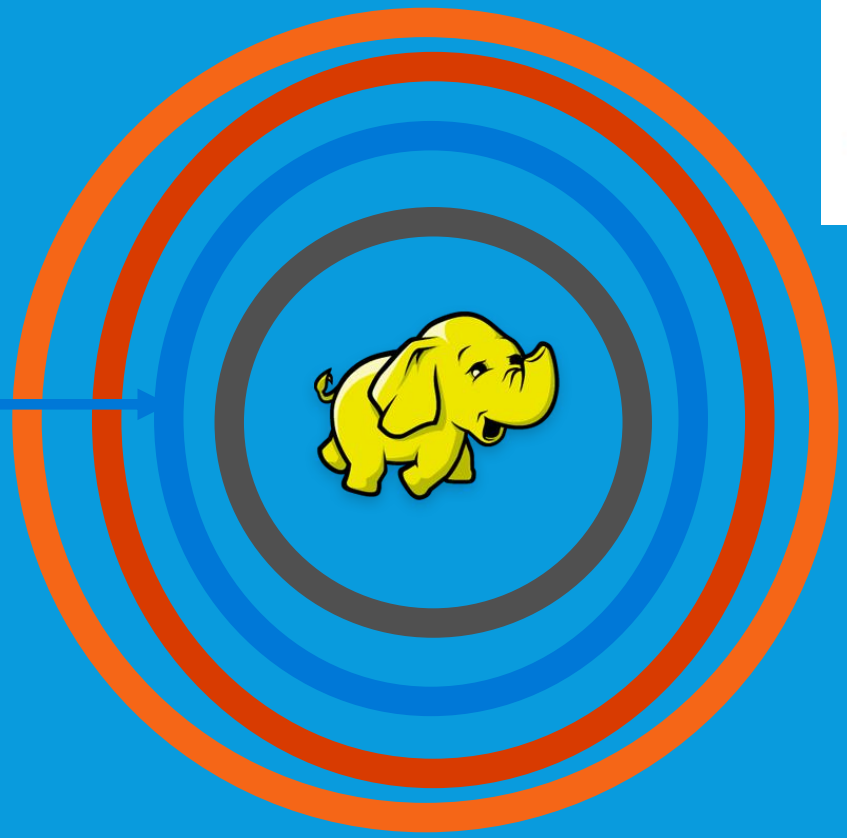
Authentication
Kerberos
Active Directory





APPLICATION AND DATA-LEVEL AUTHORIZATION

Authorization
Hive policies
HBase policies
File and Folder level ACLS



Apache Ranger

SECURE ENDPOINTS IN HDINSIGHT CLUSTER

Access to all users

- HiveServer2
- Ambari & Views
- Ranger

Access to only Cluster Admin

- SSH
- WebHCat
- Oozie

TRANSPARENT SERVER SIDE ENCRYPTION

Azure Data Lake Storage

- Public Preview
- ALWAYS ON transparent encryption
- All reads/writes are encrypted/decrypted
- Service managed keys as well as Customer managed keys

Windows Azure Storage Blob

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