Microsoft

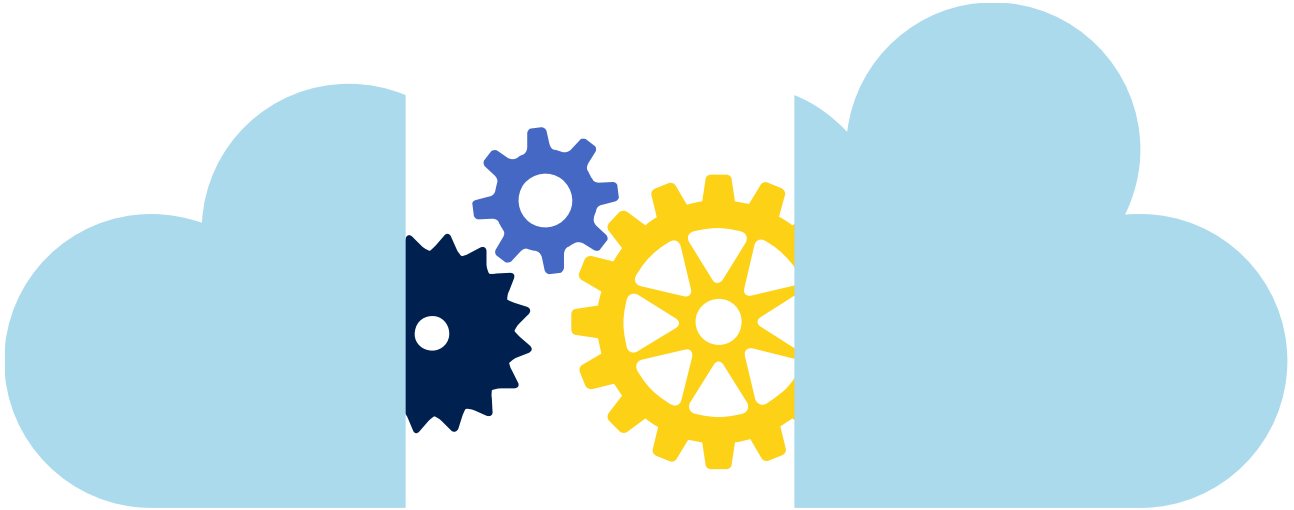
MTC Tech Summit

MTC Seattle
Hyun (hyssh@microsoft.com)

Microsoft

Agenda

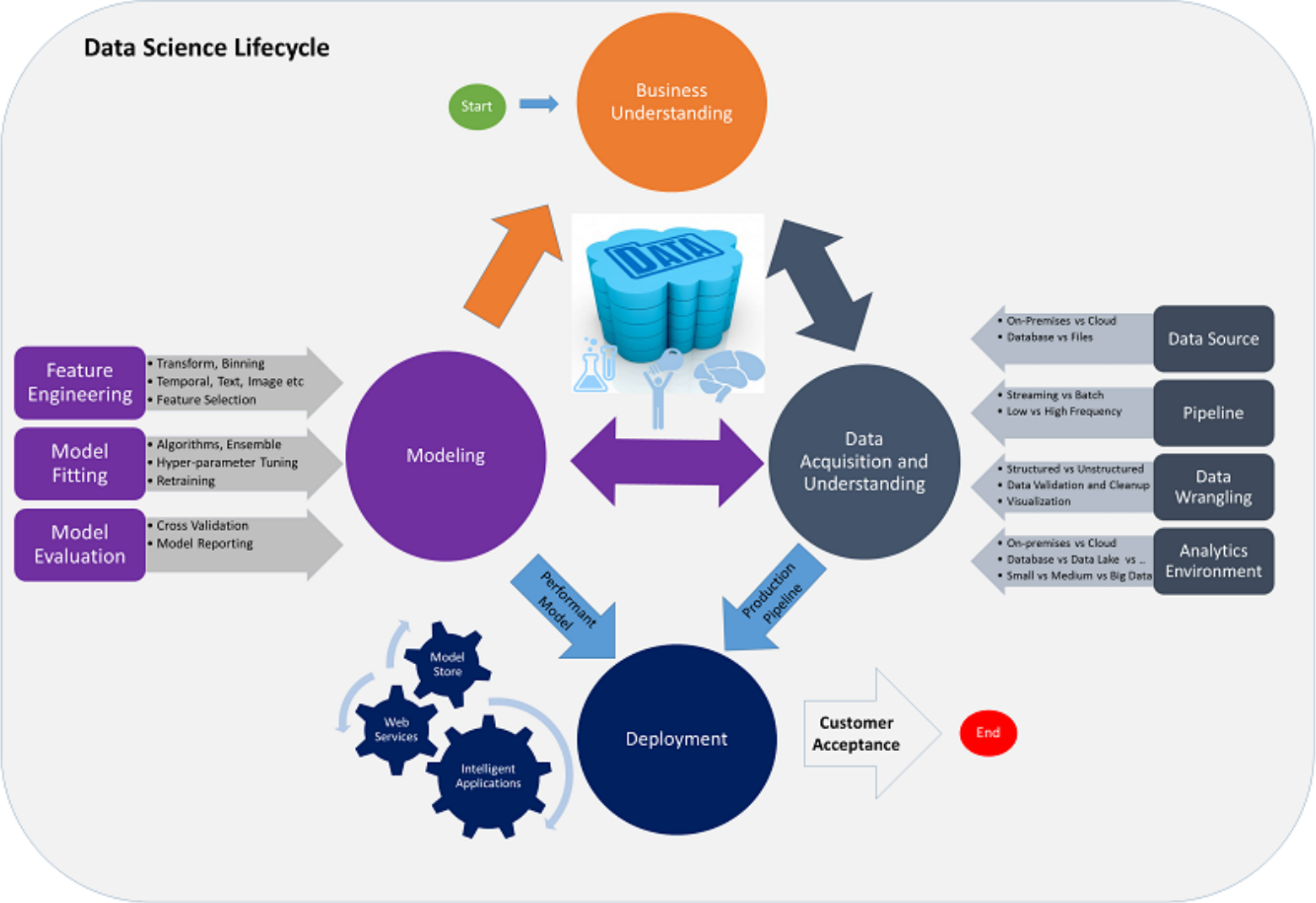
- **Data Science process – TDSP (Team Data Science Process)**
- Data Science Tools
- Hand on Labs



Data Science process – TDSP (Team Data Science Process)

Data Science Lifecycle

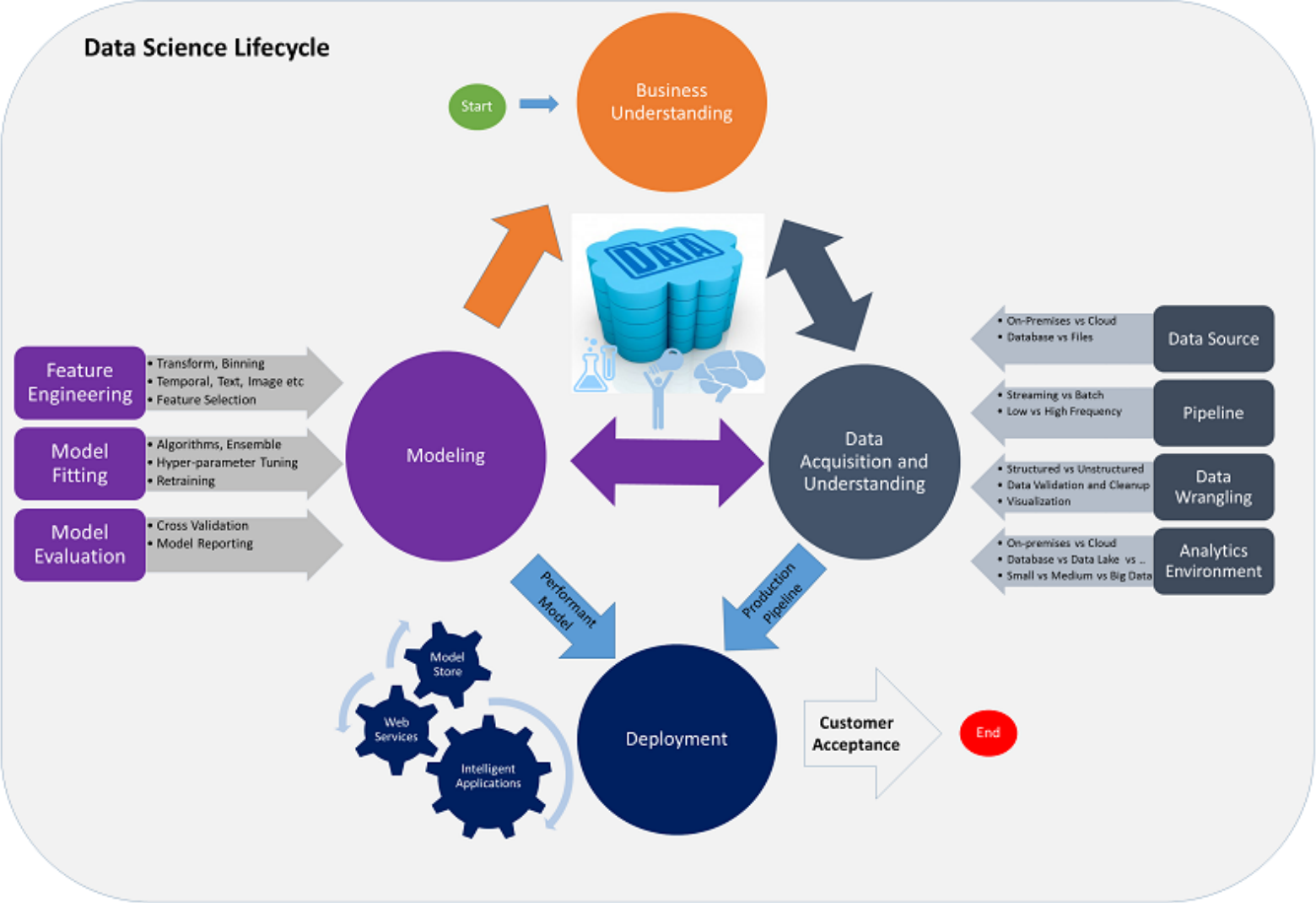
1. Business understanding
2. Data acquisition and understanding
3. Modeling
4. Deployment
5. Customer Acceptance



Data Science process – TDSP (Team Data Science Process)

Data Science Lifecycle

1. Business understanding
2. Data acquisition and understanding
3. Modeling
4. Deployment
5. Customer Acceptance



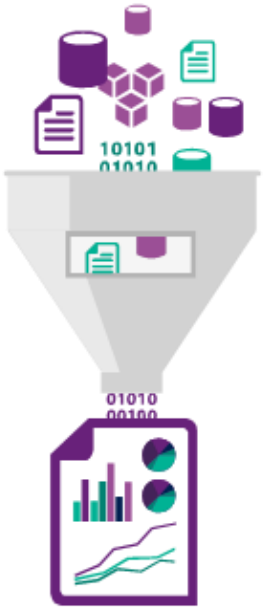
1. Business Understanding

Analyze business needs

- Define objectives
  - Identify key business variables
  - Project goal
  - Question
  - Success metrics

- Identify data sources
  - Data sources that contain known examples of answers
  - Data characteristics
  - Data quality
  - Tools and language
  - Security
    - Encryption, Audit, Access control

- Artifacts
  - Documents
  - Data sources
  - Data



1. Business Understanding

Business sample case

Wide World Importers is a company that imports and distributes products in multiple countries around the globe.

With several thousand employees, Information Technology is at the heart of our business operations, and has a significant cost.

Since we handle materials in multiple countries, we have a lot of private data, financial information, and other targets which have a high security profile. We are concerned with both external and internal attacks. In addition, many of our employees work in remote locations, some on ships and other challenging environments.

All of our IT systems have been modernized, and we're taking in a significant amount of semi-structured data from computing devices – most of it real-time. After talking with our IT leadership, we need a way to determine anomalies within the data streams we get, and have a way to observe the anomalies in a dashboard so that we can respond to outages, threats, and changes quickly.

1. Business Understanding

Design statements

Wide World Importers is a company that imports and distributes products in multiple countries around the globe.

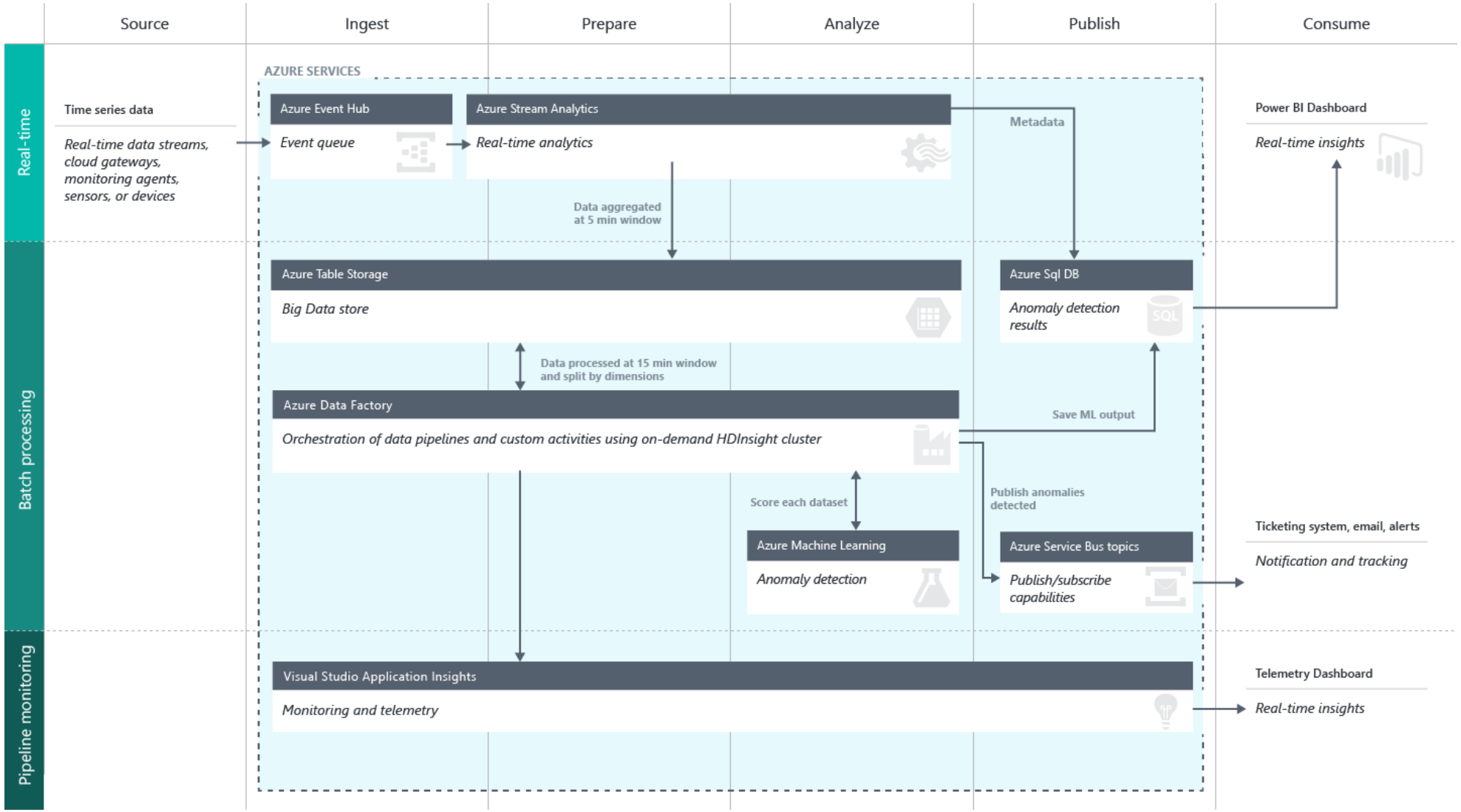
With several thousand employees, Information Technology is at the heart of our business operations, and has a significant cost.

Since we handle materials in multiple countries, we have a lot of private data, financial information, and other targets which have a high security profile. We are concerned with both external and internal attacks. In addition, many of our employees work in remote locations, some on ships and other challenging environments.

All of our IT systems have been modernized, and we're taking in a significant amount of semi-structured data from computing devices – most of it real-time. After talking with our IT leadership, we need a way to determine anomalies within the data streams we get, and have a way to observe the anomalies in a dashboard so that we can respond to outages, threats, and changes quickly.

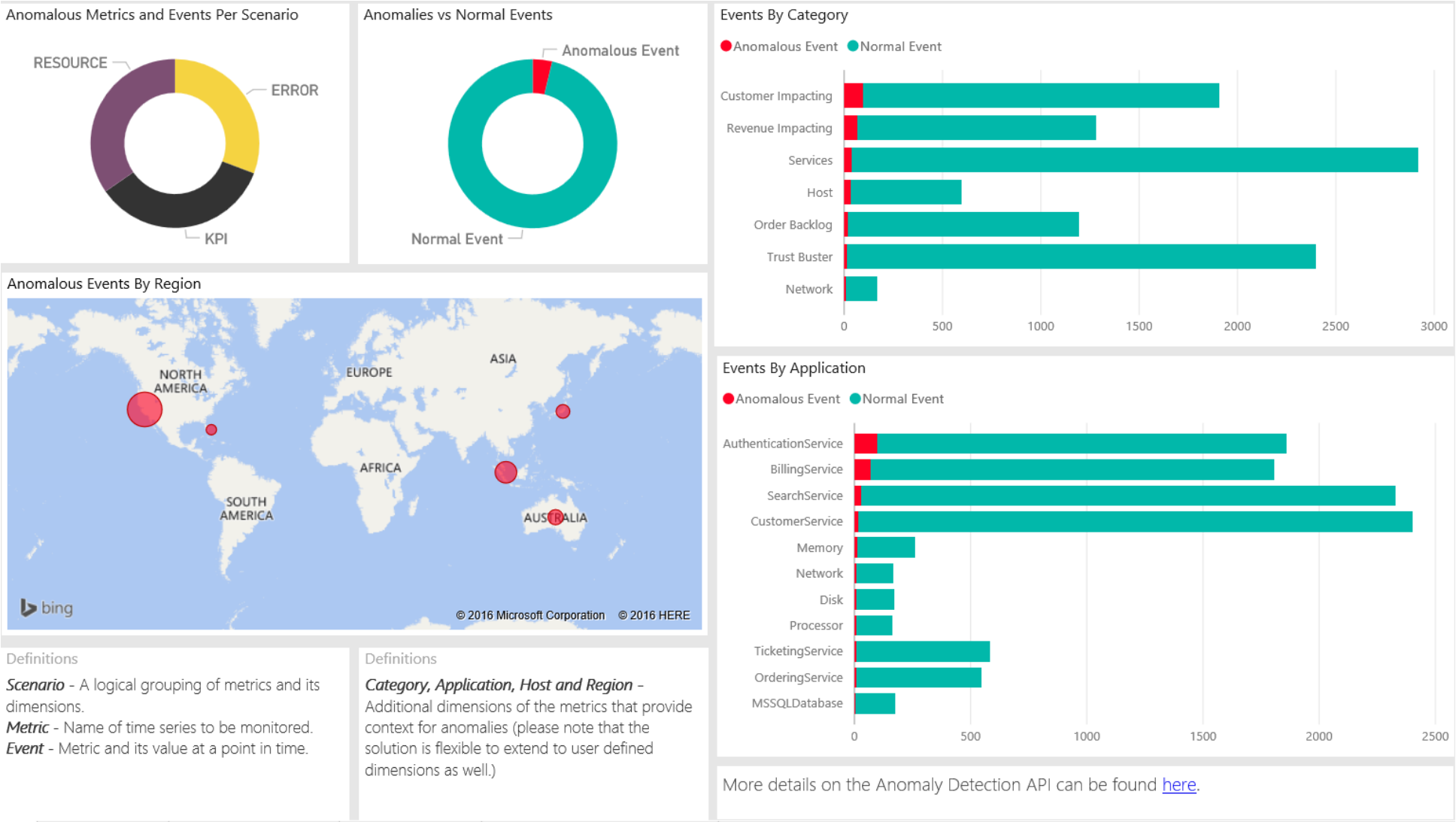
1. Business Understanding

Solution Diagram



1. Business Understanding

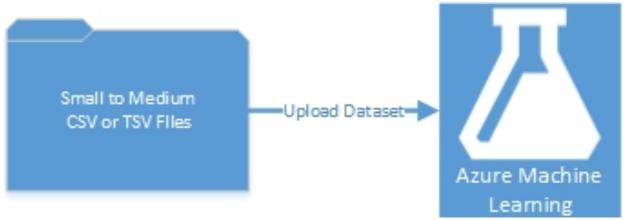
Report



1. Business Understanding

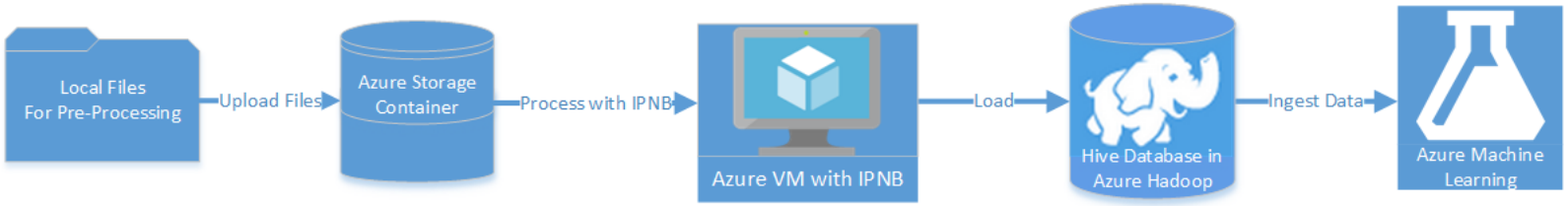
Solution Diagram

- Scenario #1: Small to medium tabular dataset in a local files



- Scenario #2: Small to medium dataset of local files that require processing

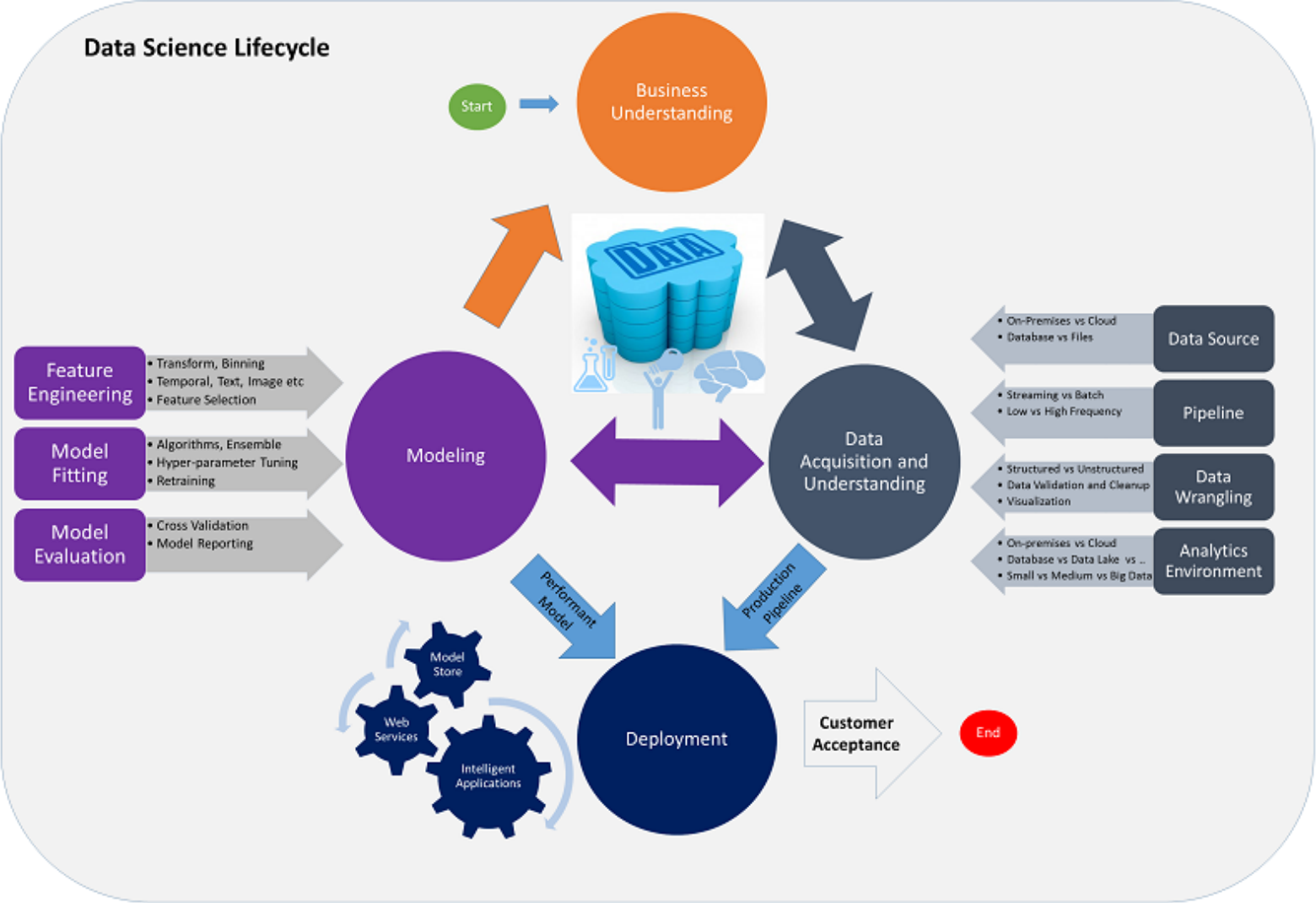
- Scenario #3: Big data in local files, target Hive database in Azure HDInsight Hadoop clusters



Data Science process – TDSP (Team Data Science Process)

Data Science Lifecycle

1. Business understanding
2. Data acquisition and understanding
3. Modeling
4. Deployment
5. Customer Acceptance



2. Data Acquisition and Understanding

Acquire and understand data

- Ingest the data
  - Move the data from source to target location

- Explore the data
  - Data cleaning
    - Incomplete
    - Noisy
    - Inconsistent
  - Audit the quality of data
  - Have better understand the patterns
  - Iterative works

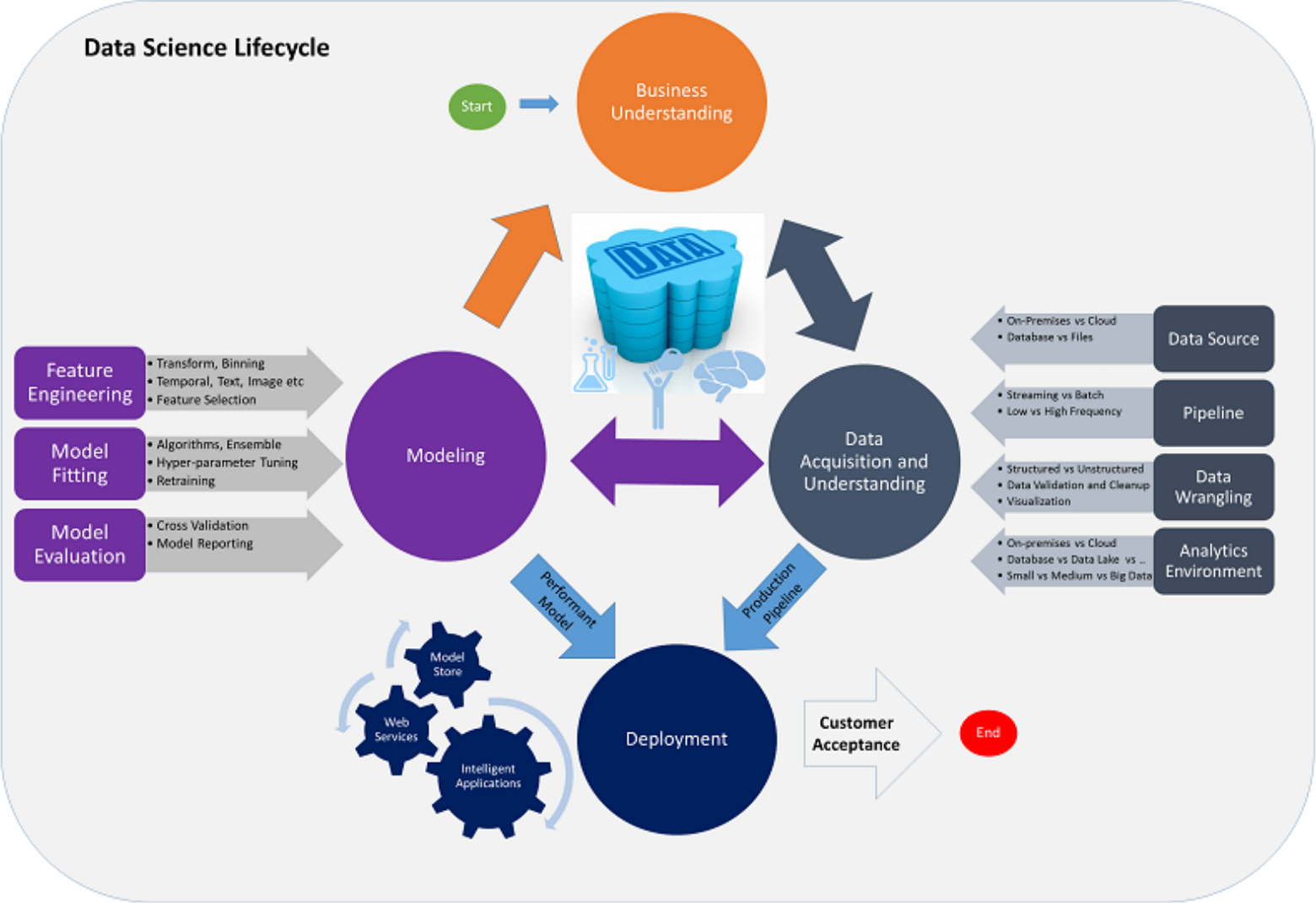
- Set up a data pipeline
  - Score new data
  - Refresh the data regularly
  - Pipeline may be batch-based or a streaming or hybrid

- Artifacts
  - Data quality report
  - Solution architecture
  - Checkpoint decision

Data Science process – TDSP (Team Data Science Process)

Data Science Lifecycle

1. Business understanding
2. Data acquisition and understanding
3. Modeling
4. Deployment
5. Customer Acceptance



3. Modeling

Develop models

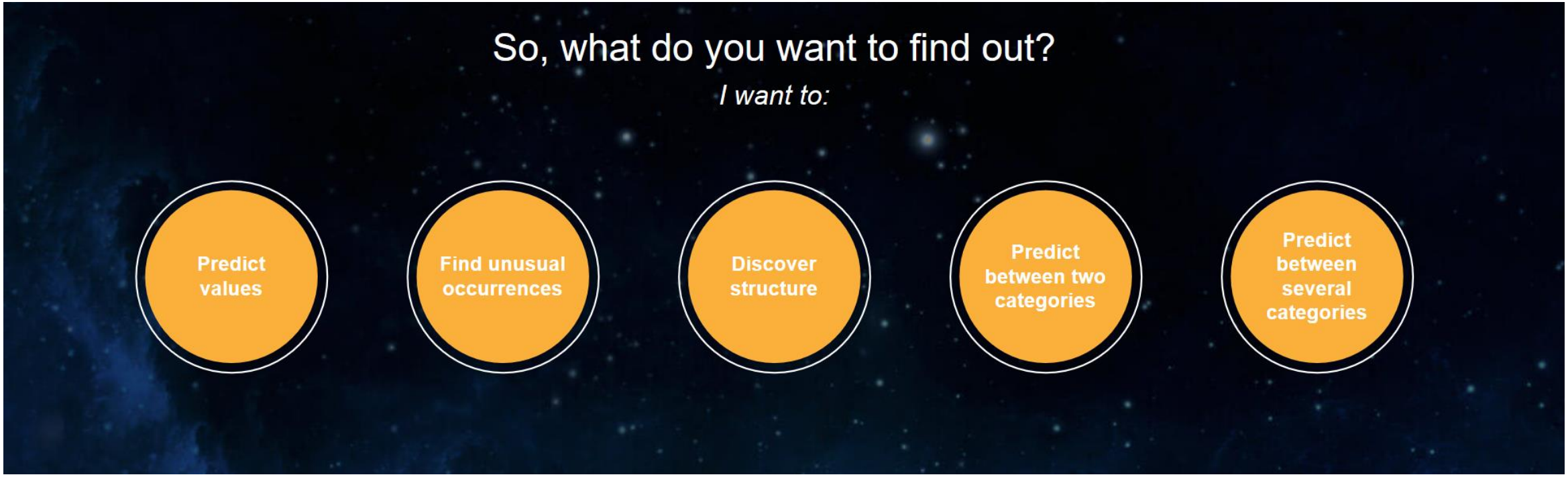
- Goals
  - Optimal data features for the machine learning model
  - Build informative machine learning model that predicts the target most accurately
  - Experiment for machine learning model that is suitable for production

- How to do it
  - Feature engineering
    - Aggregate and transform the raw variables to create the features
    - Requires a creative combination of domain expertise and insights
  - Model training
    - How to choose algorithms
    - Split, Build, Evaluate, Determine the best solution

- Artifacts
  - Feature sets
  - Modeling report
  - Checkpoint decision

3. Modeling

How to choose algorithms

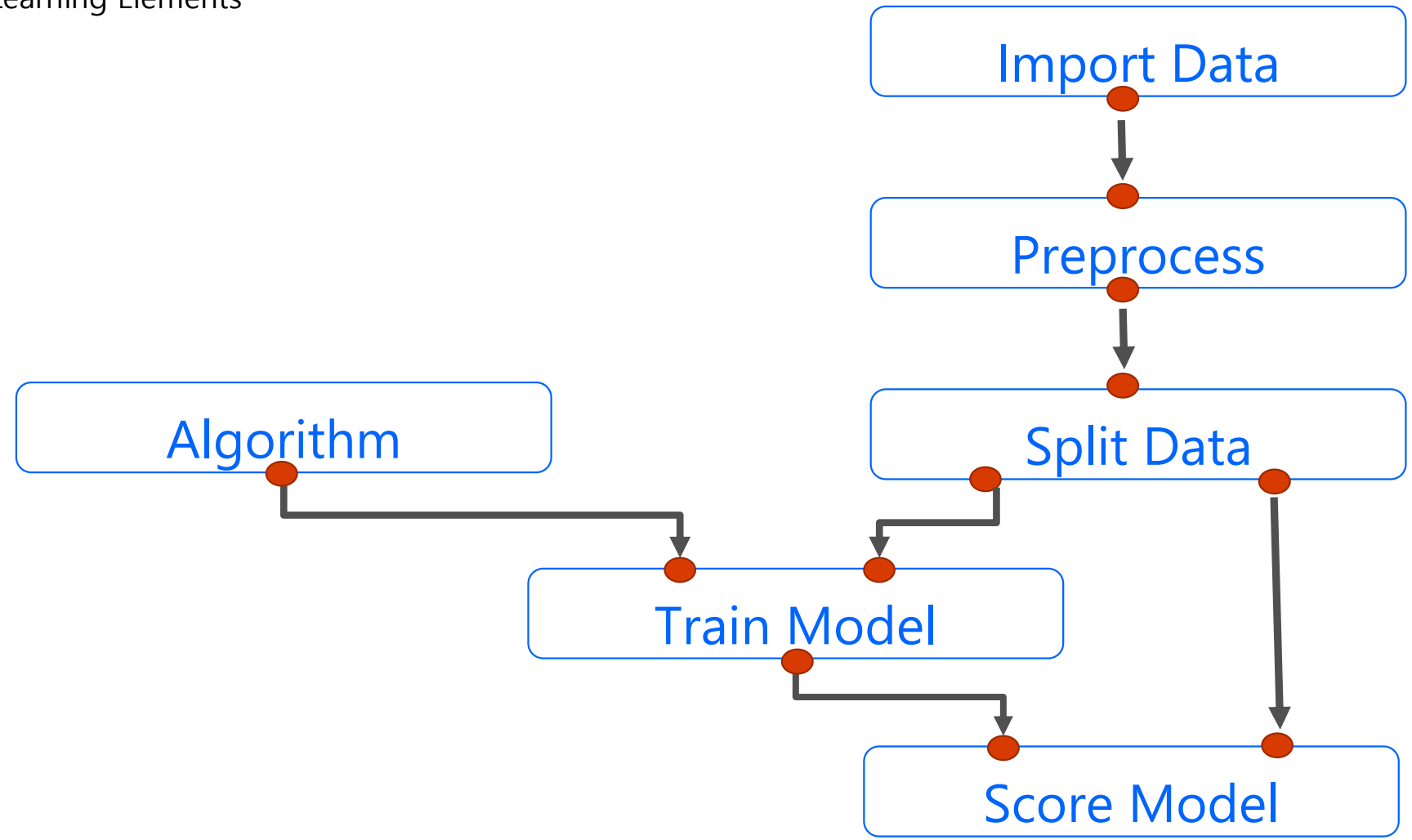


- https://azuremlsimpleds.azurewebsites.net/simpleds/

3. Modeling

Model Training

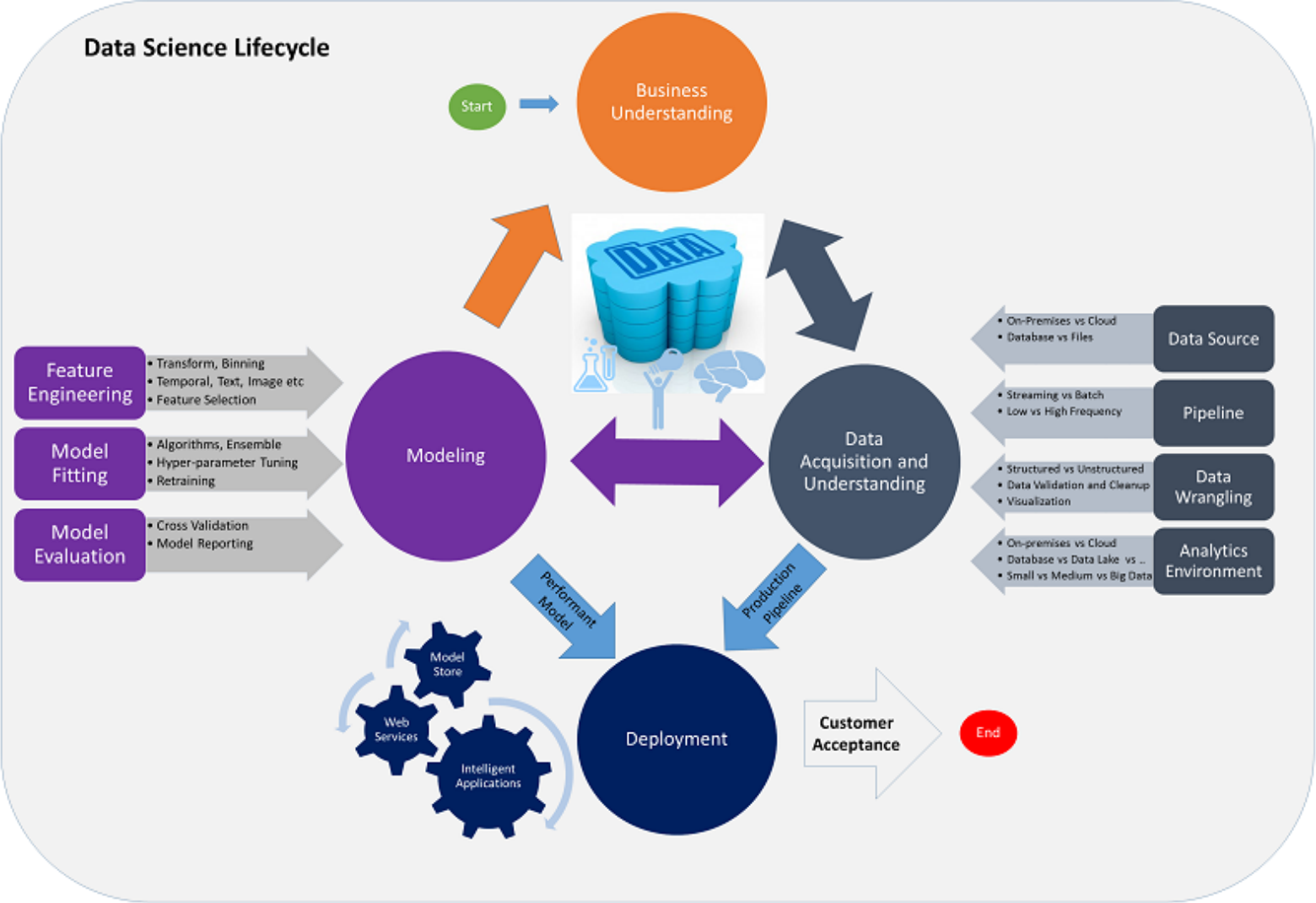
- Basic Machine Learning Elements



Data Science process – TDSP (Team Data Science Process)

Data Science Lifecycle

1. Business understanding
2. Data acquisition and understanding
3. Modeling
4. Deployment
5. Customer Acceptance



4. Deployment

Operationalize models

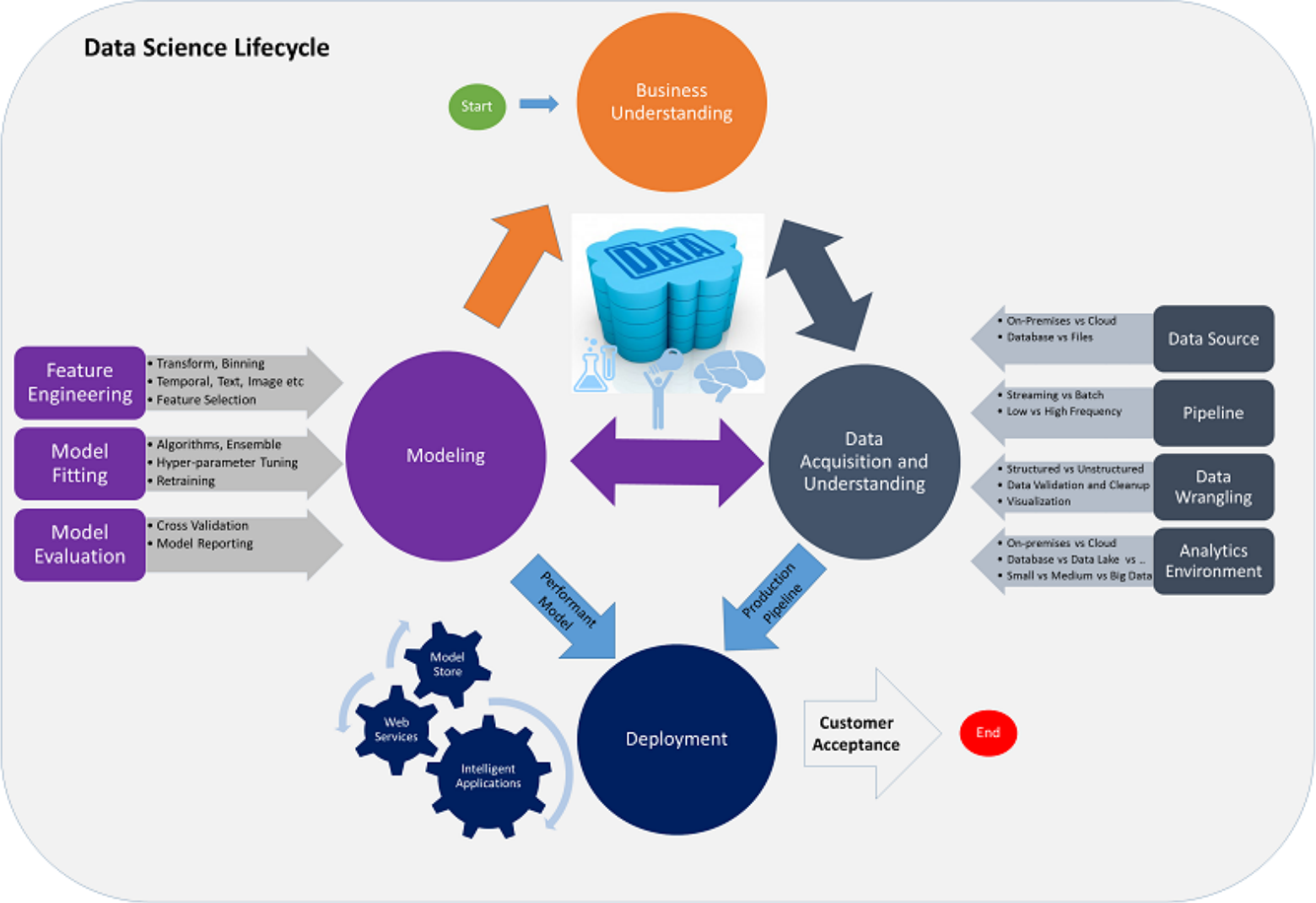
- Operationalize the model
  - Apply trained model to application for real-time or batch basis predictions
  - Expose the model as API interface to consume it easy

- Artifacts
  - Status dashboard of system health and key metrics
  - Final modeling report with deployment details
  - Final solution architecture document

Data Science process – TDSP (Team Data Science Process)

Data Science Lifecycle

1. Business understanding
2. Data acquisition and understanding
3. Modeling
4. Deployment
5. Customer Acceptance



5. Customer Acceptance

Finalize the project deliverables

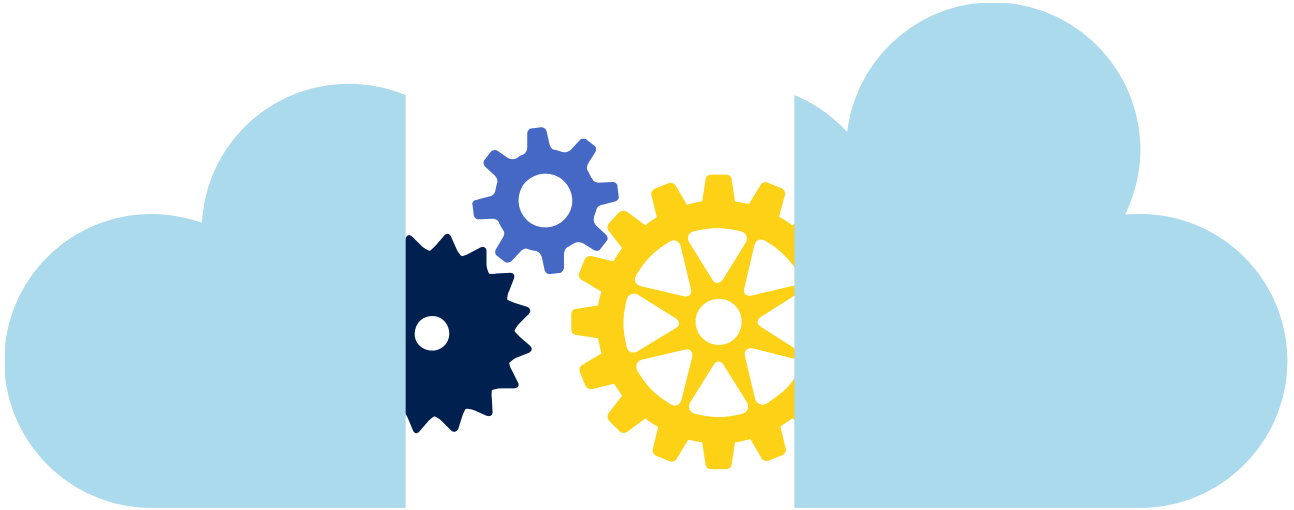
- System validation
  - Confirm the deployed model and pipeline are meeting customer needs
  - Monitor systems

- Project hand-off

- Artifacts
  - Project final report

Agenda

- Data Science process – TDSP (Team Data Science Process)
- **Data Science Tools**
- Hands on Lab



Data Science Tools

TDSP

Business Understanding	•Define Objectives •Identify Data Sources
Data Acquisition and Understanding	•Ingest Data •Explore Data •Update Data
Modeling	•Feature Selection •Create and Train Model
Deployment	•Operationalize
Customer Acceptance	•Testing and Validation •Handoff •Re-train and re-score

Azure Services

Information Management	• **Data Catalog** • **Data Factory** • **Event Hubs**
Big Data	• **Azure Storage** • **Data Lake** • **SQL Data Warehouse** • **Cosmos DB**
Intelligence and Advanced Analytics	• **Cortana, Bot Service, Cognitive Framework** • **Machine Learning** • **HDInsight** • **Data Lake** • **Stream Analytics** • **Analysis Services**
Visualization	• **Power BI** • **R**
Solutions	• **Templates and Gallery**

Data Science Tools

TDSP

Business Understanding	•Define Objectives •Identify Data Sources
Data Acquisition and Understanding	•Ingest Data •Explore Data •Update Data
Modeling	•Feature Selection •Create and Train Model
Deployment	•Operationalize
Customer Acceptance	•Testing and Validation •Handoff •Re-train and re-score

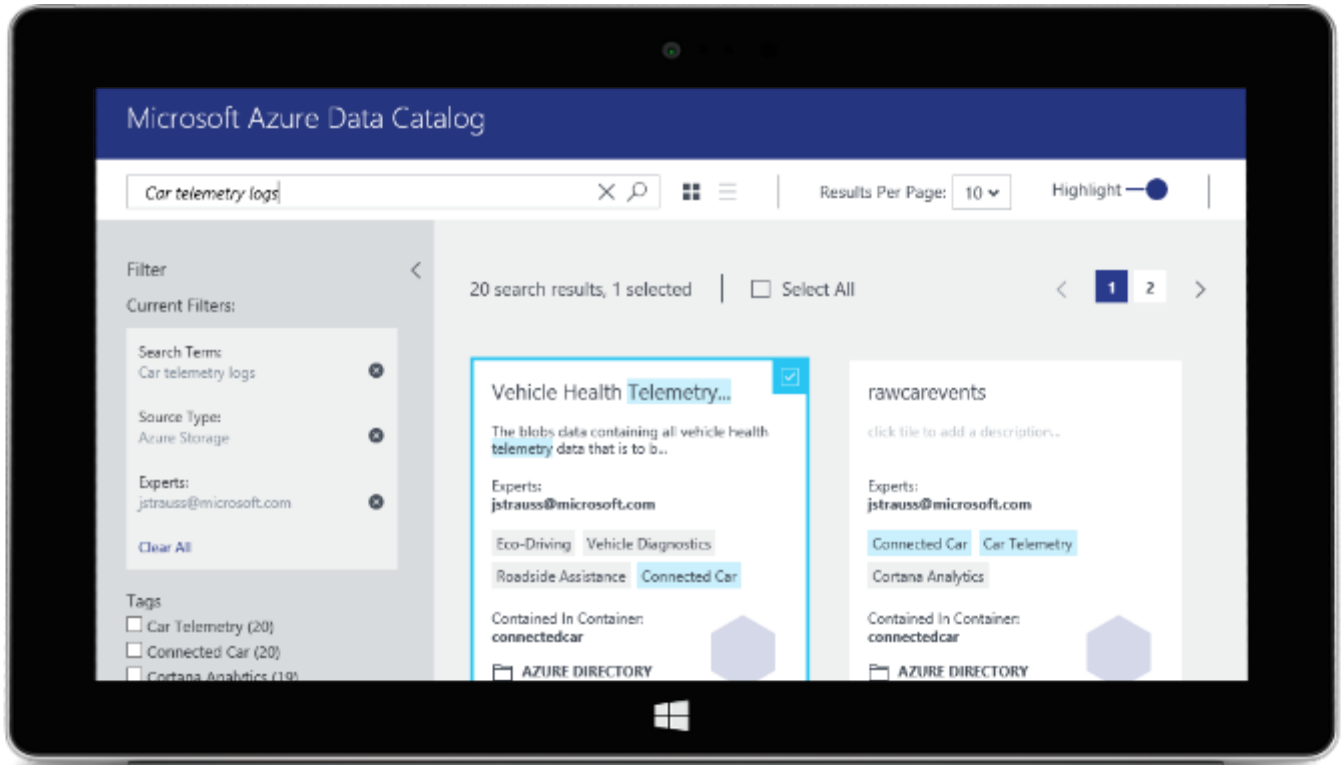
Azure Services

Information Management	• **Data Catalog** • **Data Factory** • **Event Hubs**
Big Data	• **Azure Storage** • **Data Lake** • **SQL Data Warehouse** • **Cosmos DB**
Intelligence and Advanced Analytics	• **Cortana, Bot Service, Cognitive Framework** • **Machine Learning** • **HDInsight** • **Data Lake** • **Stream Analytics** • **Analysis Services**
Visualization	• **Power BI** • **R**
Solutions	• **Templates and Gallery**

1. Business Understanding

Analyze business needs

- Find data sources using Azure Data Catalog



- Spend less time looking for data, and more time getting value from it
- Register enterprise data sources, discover data assets and unlock their potential, and capture tribal knowledge to make data understandable
- Bridge the gap between IT and the business, allowing everyone to co ntribute their insights, tags, and descriptions
- Intuitive search and filtering to understand the data sources and t heir purpose
- Let your data live where you want; connect using tools you choos e
- Integrate into existing tools and processes with open REST APIs

Data Catalog

https://www.azuredatacatalog.com

Data Science Tools

TDSP

Business Understanding	•Define Objectives •Identify Data Sources
Data Acquisition and Understanding	•Ingest Data •Explore Data •Update Data
Modeling	•Feature Selection •Create and Train Model
Deployment	•Operationalize
Customer Acceptance	•Testing and Validation •Handoff •Re-train and re-score

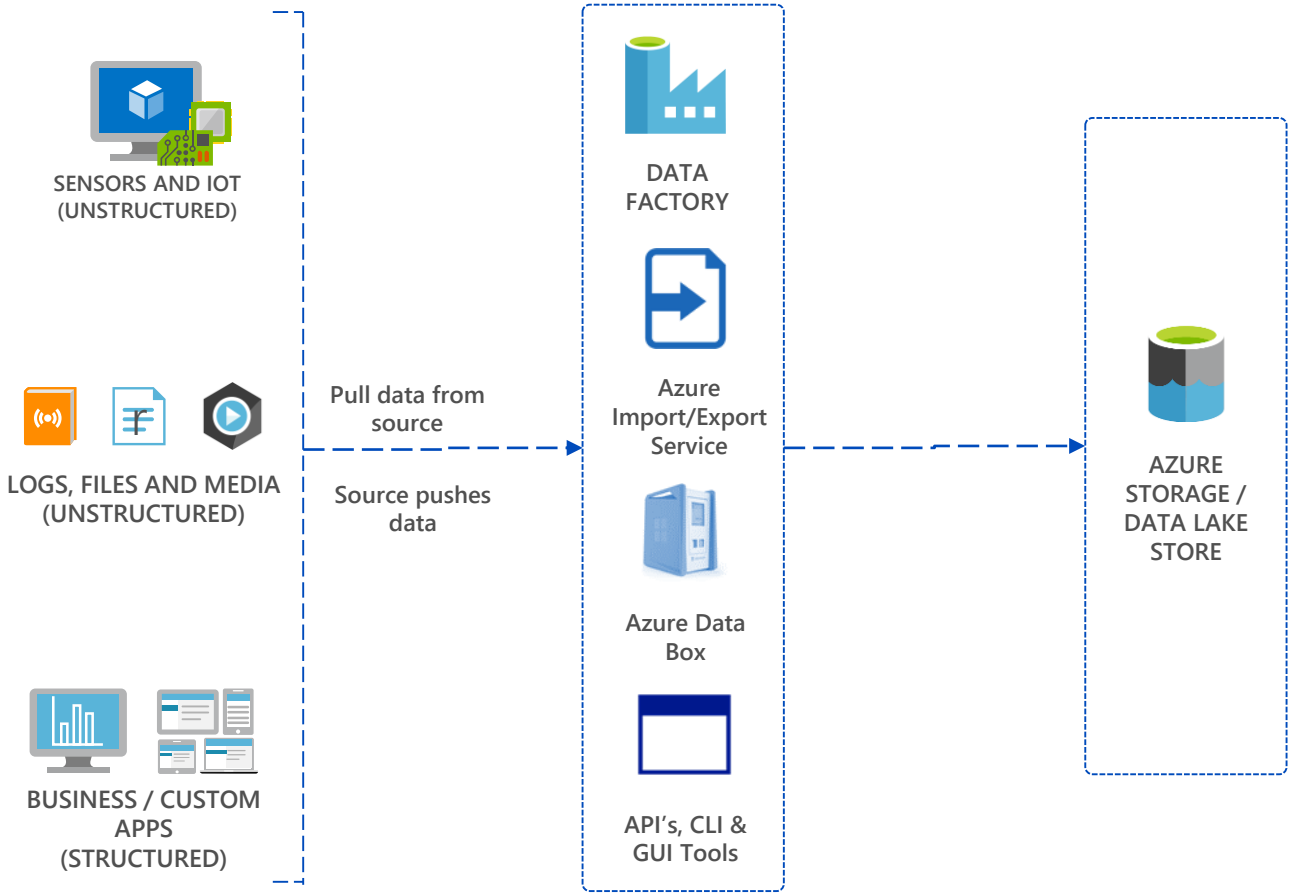
Azure Services

Information Management	• Data Catalog • Data Factory • Event Hubs
Big Data	• Azure Storage • Data Lake • SQL Data Warehouse • Cosmos DB
Intelligence and Advanced Analytics	• Cortana, Bot Service, Cognitive Framework • Machine Learning • HDInsight • Data Lake • Stream Analytics • Analysis Services
Visualization	• Power BI • R
Solutions	• Templates and Gallery

2. Data Acquisition and Understanding

Ingest data

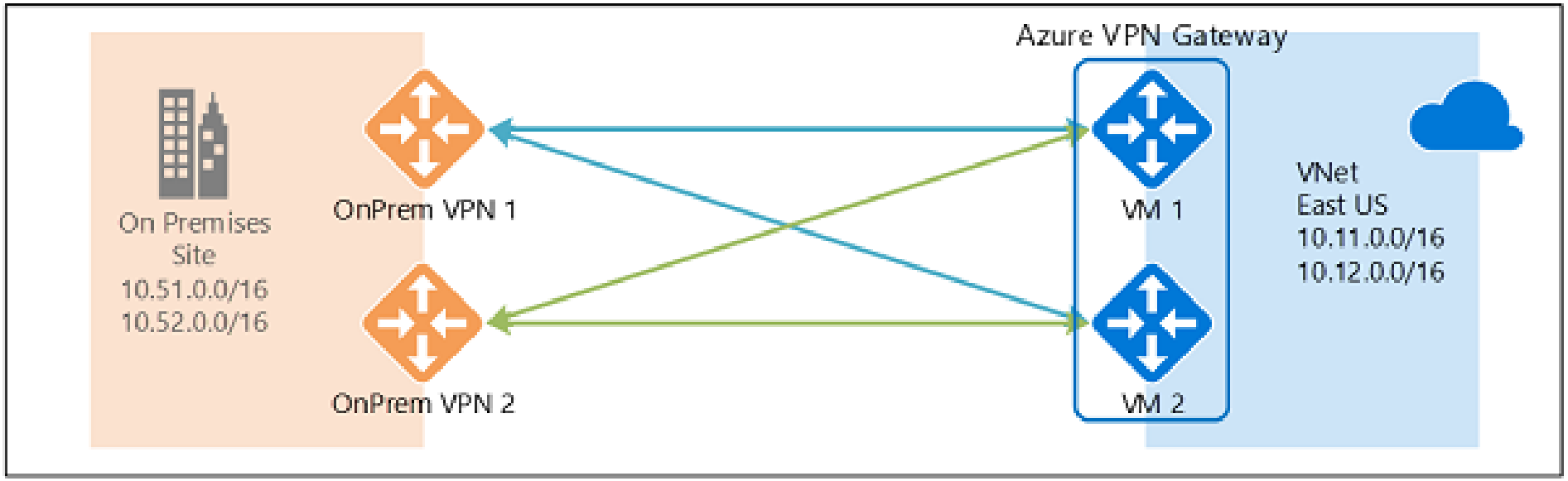
Store data



2. Data Acquisition and Understanding

Move data via network

- Connect on-premises to <anything>
  - VPN Gateway
    - Send network traffic from virtual networks to on-prem locations
    - Send network traffic between virtual networks within Azure
    - Site-to-site vs. Point-to-site
    - You can connect multiple on-prem locations to a virtual network (Multi-site)
  - ExpressRoute can directly connect your WAN to Azure

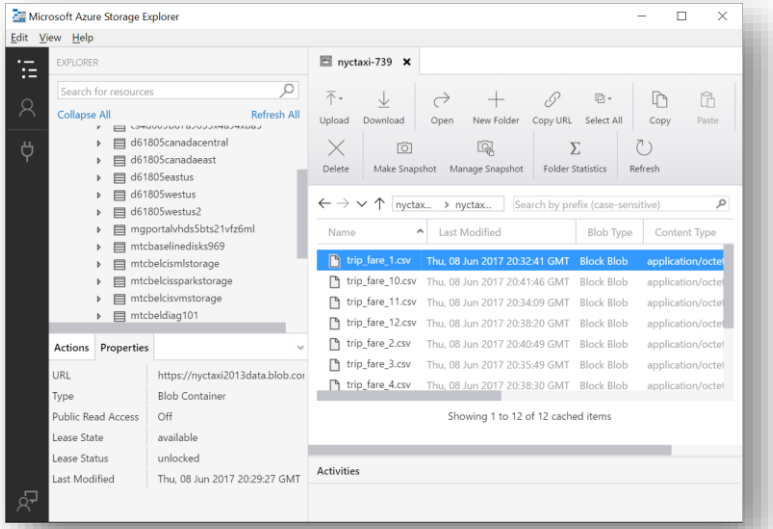


2. Data Acquisition and Understanding

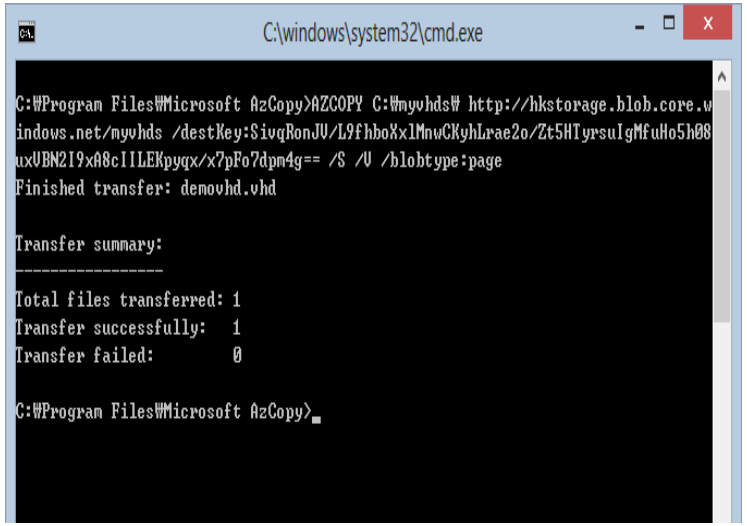
Acquire and understand data

- Azure portal
- PowerShell
- Azure Data Factory
- Azure Event Hubs
- Azure storage SDKs (.NET, Node.js, python, C++, etc.)
- AzCopy (blob, file, and table only)
- Import/Export service

[Azure Portal]



[Azure Storage Explorer]

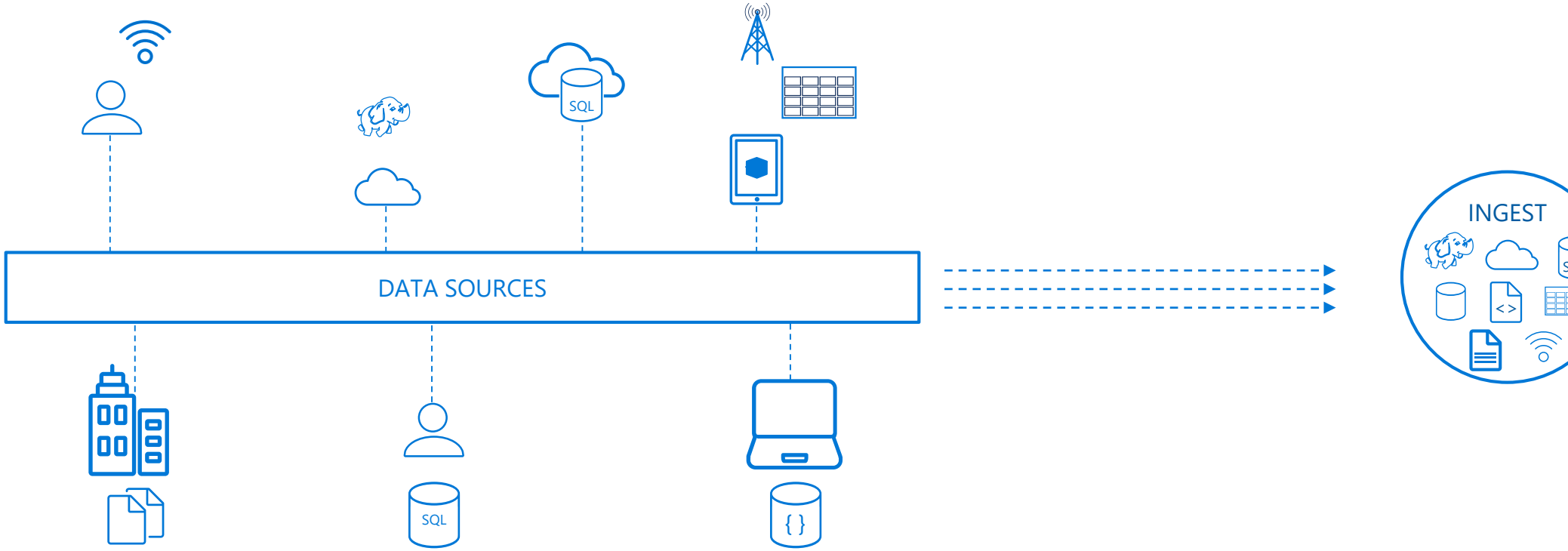


[AZcopy]

2. Data Acquisition and Understanding

Ingest Data

- Azure Data Factory



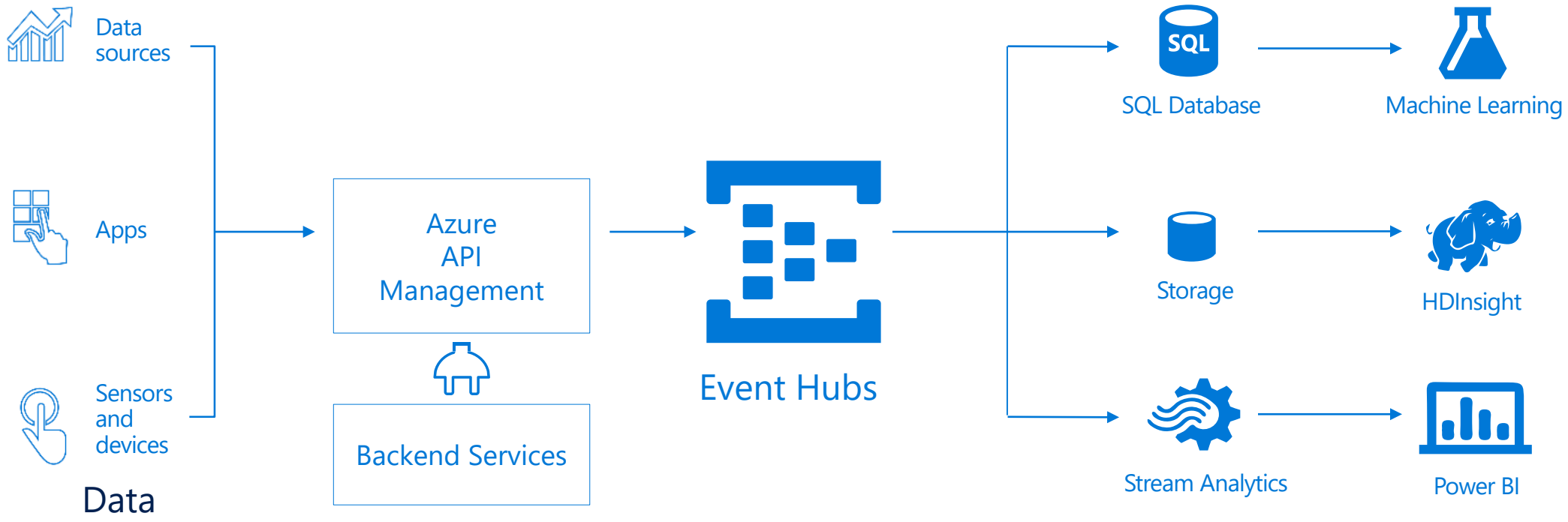
- Create, schedule, orchestrate, and manage data pipelines
- Visualize data lineage
- Connect to on-premises and cloud data sources
- Monitor data pipeline health

- Automate cloud resource management
- Move relational data for Hadoop processing
- Transform with Hive, Pig, or custom code

2. Data Acquisition and Understanding

Ingest Real-time data

- Event Hub



- Log millions of events per second in near real time
- Connect devices using flexible authorization and throttling
- Use time-based event buffering
- Get a managed service with elastic scale

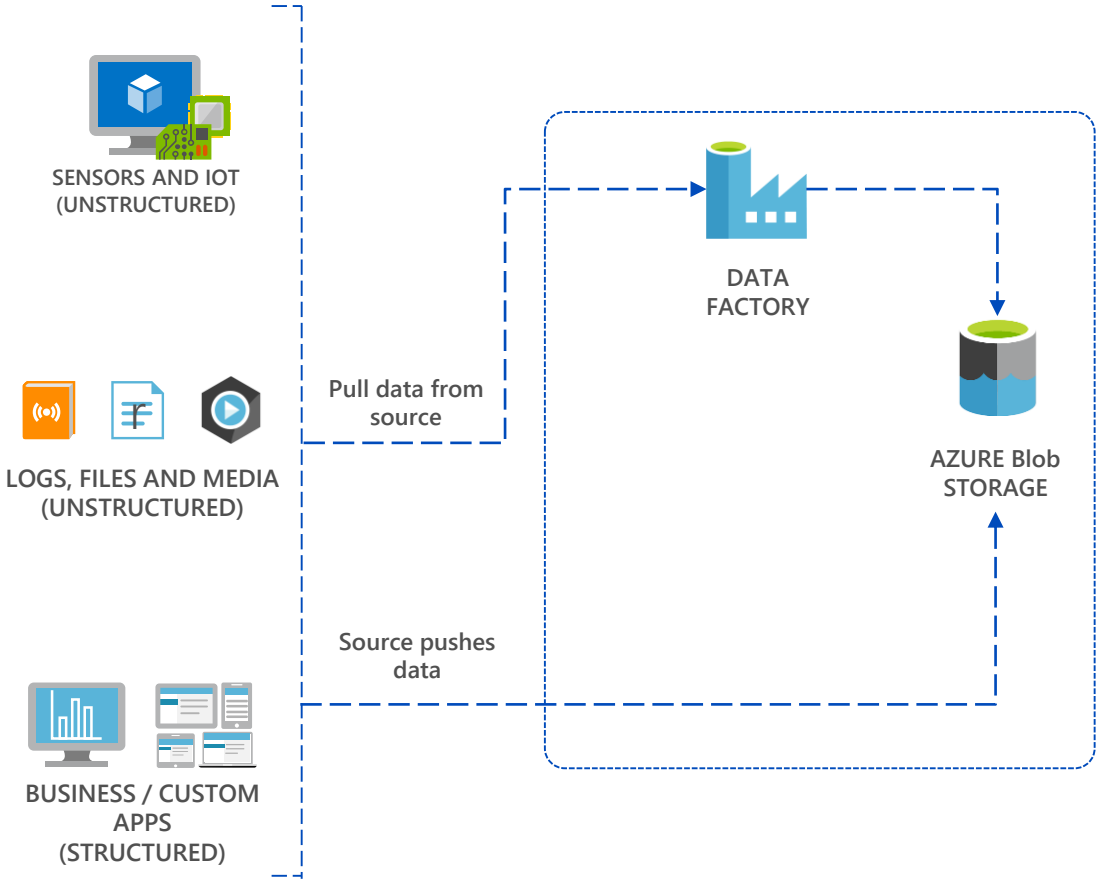
- Get a managed service with elastic scale
- Reach a broad set of platforms using native client libraries
- Pluggable adapters for other cloud services

2. Data Acquisition and Understanding

Ingest data

Store data

Understand data



2. Data Acquisition and Understanding

Store data

Azure Blob Storage – A highly scalable object storage for unstructured data

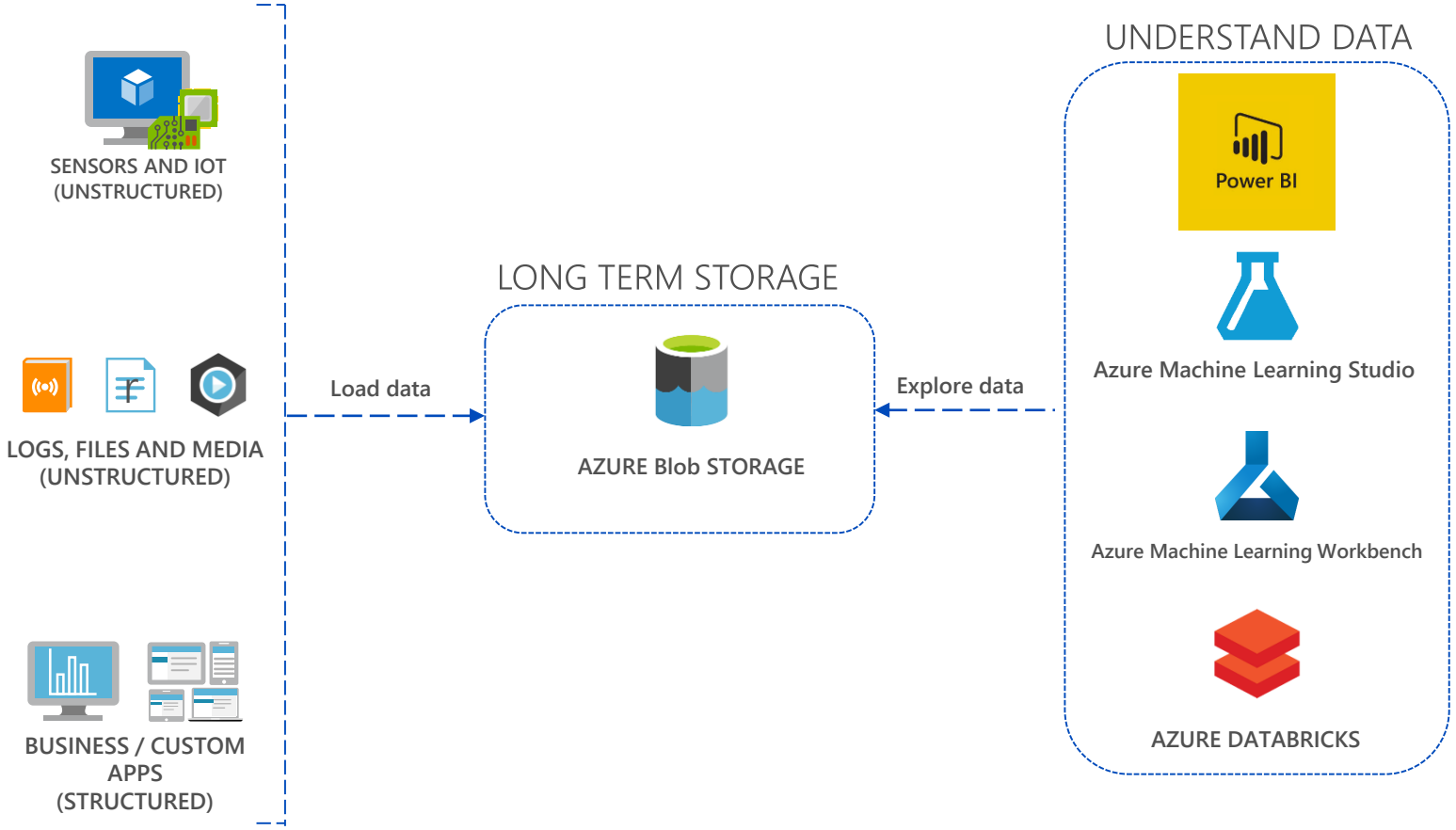
- Serverless Azure Service.
- Automatically scales as more data is uploaded.
- Can store billions of objects.
- Can store Images, Videos, Audio, Documents etc.
- Three types of Blobs: Block, Append and Page. Blobs are mutable.
- Four Replication Options: LRS, GRS, ZRS and RA-GRS
- Three storage tiers – Hot, Cool and Archive. Object can move between tiers.
- Strongly consistent
- SLA: 99.9 uptime and 99.99% for reads with RA-GRA (details)
- Monitoring via Azure Monitor
- Data encrypted at rest and in motion

2. Data Acquisition and Understanding

Store data

Understand data

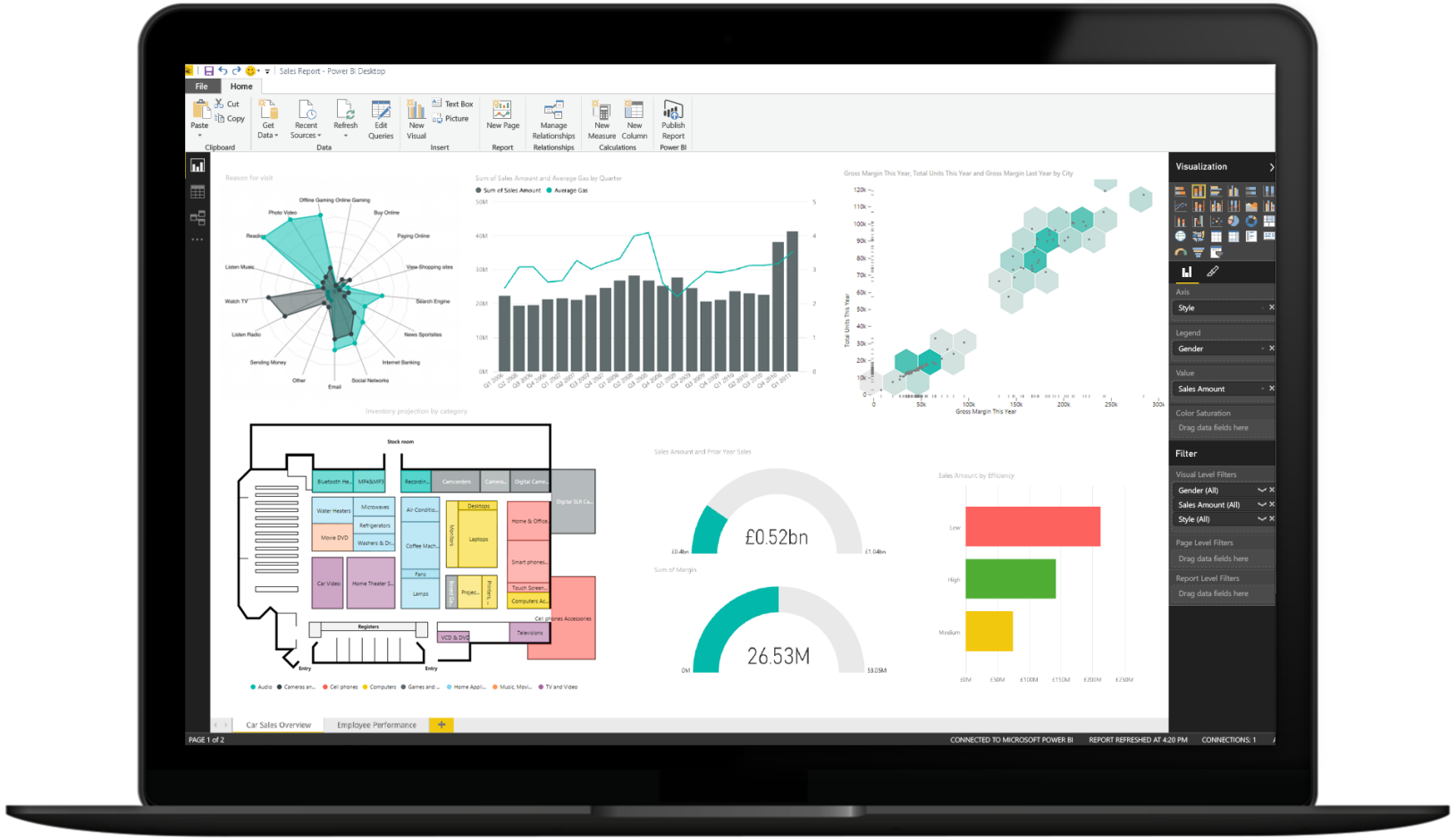
Prepare data



2. Data Acquisition and Understanding

Tools to understand data

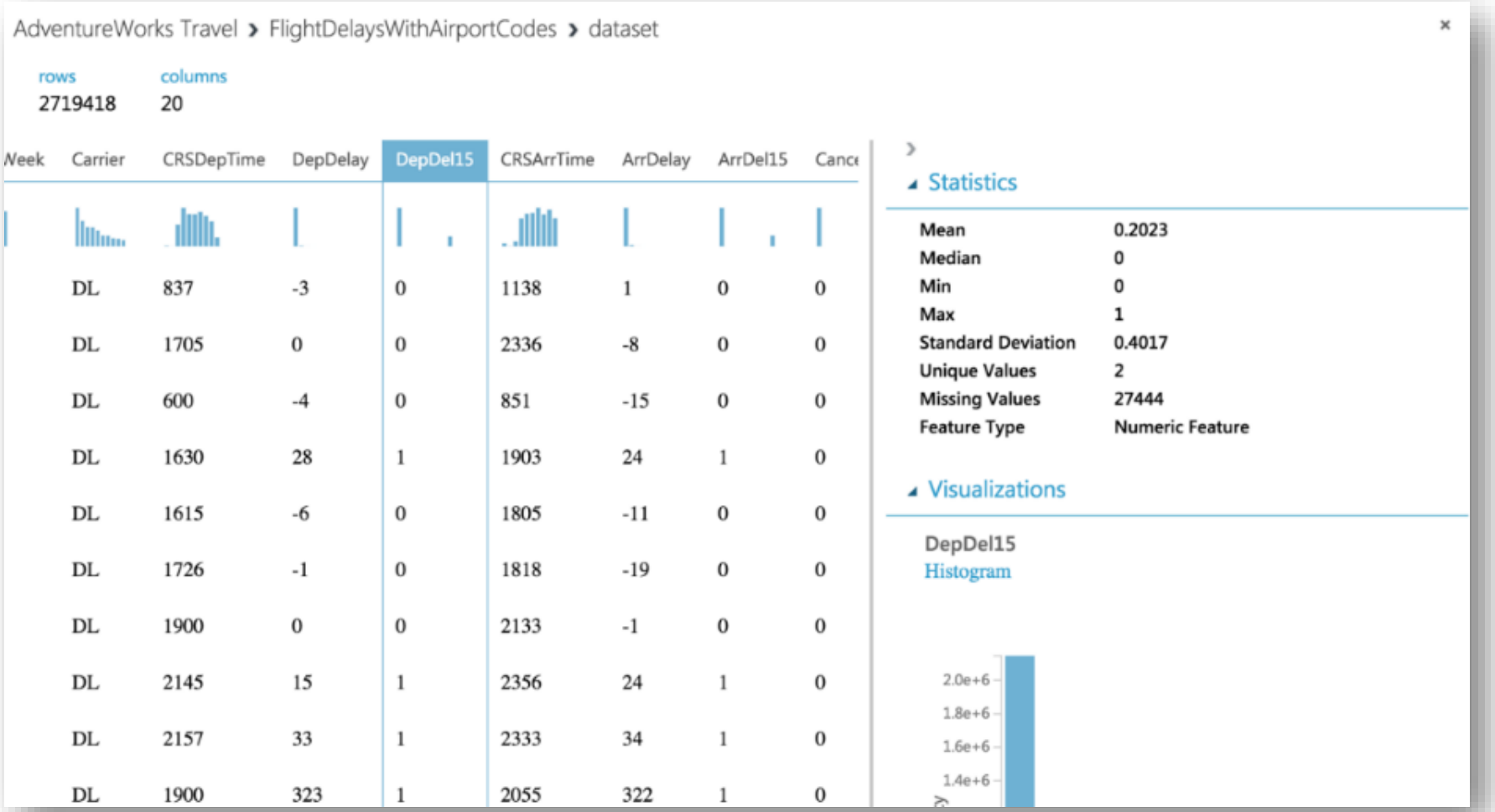
- Excel/PowerBI



2. Data Acquisition and Understanding

Tools to understand data

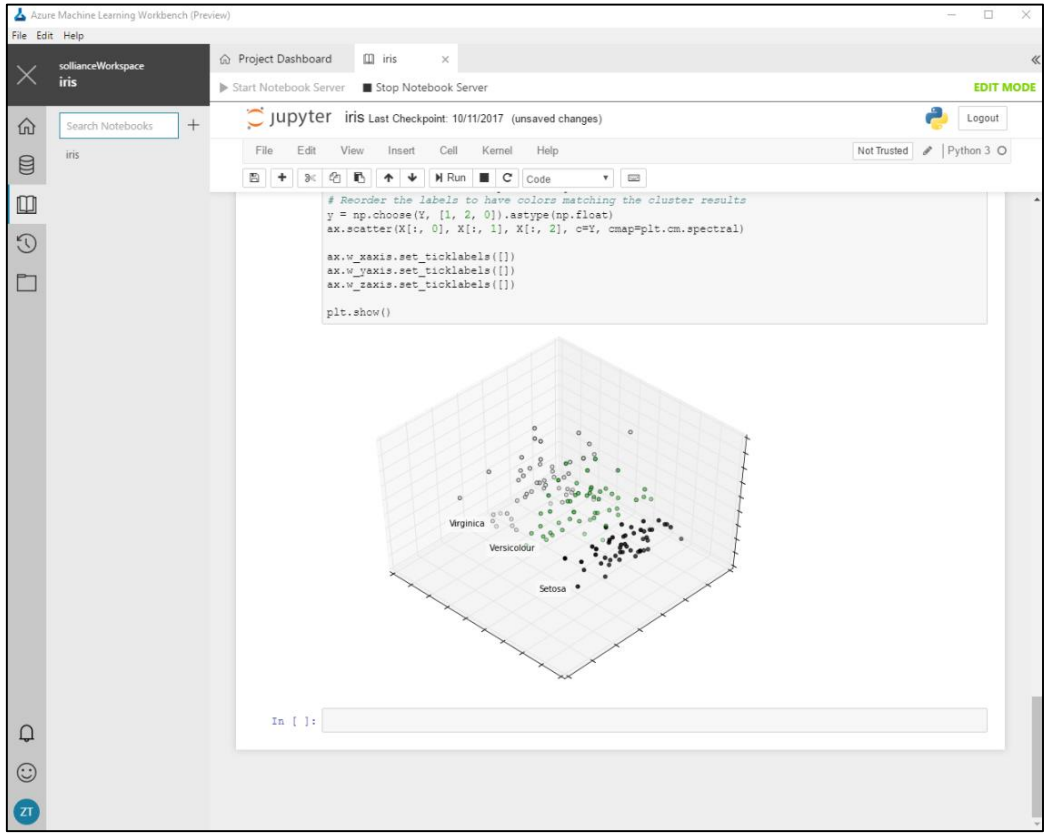
- Azure Machine Learning Studio



2. Data Acquisition and Understanding

Tools to understand data

- Azure Machine Learning Workbench
  - Desktop application plus command-line tools
  - Supports entire data science life cycle:
    - Data ingestion and preparation
    - Model development and experiment management
    - Model deployment in various target environments
  - Integrated with Azure Machine Learning services



2. Data Acquisition and Understanding

Tools to understand data

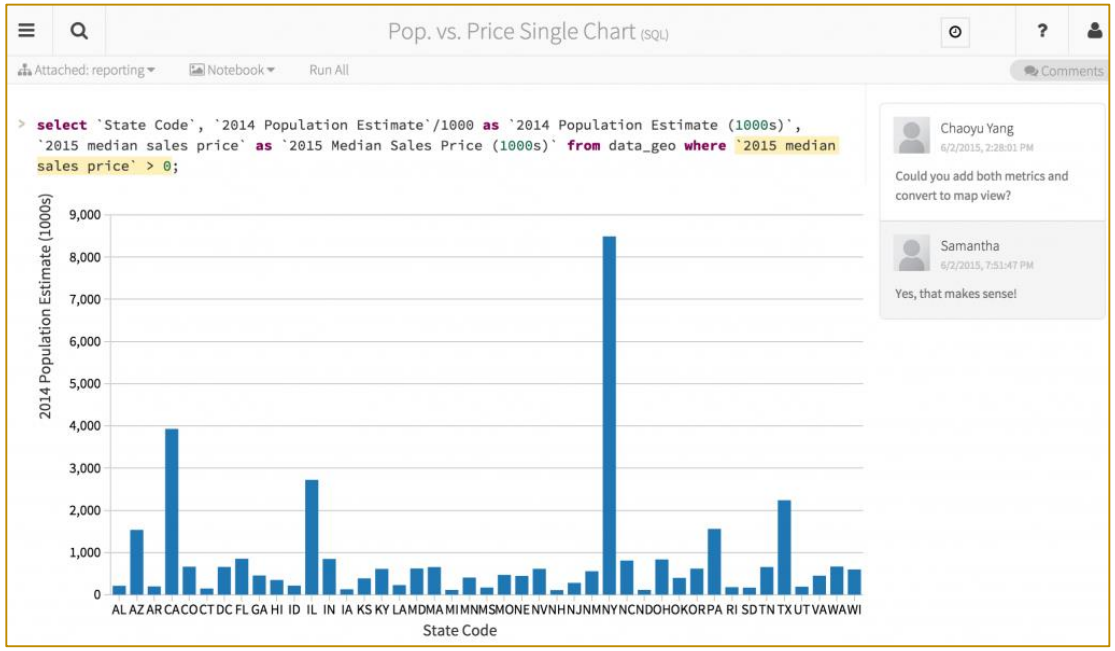
- Azure Databricks

- Azure Databricks is a **first party** service on Azure.
  - Unlike with other clouds, it is not an Azure Marketplace or a 3rd party hosted service.
- Azure Databricks is integrated seamlessly with Azure services:
  - Azure Portal: Service an be launched directly from Azure Portal
  - Azure Storage Services: Directly access data in Azure Blob Storage and Azure Data Lake Store
  - Azure Active Directory: For user authentication, eliminating the need to maintain two separate sets of uses in Databricks and Azure.
  - Azure SQL DW and Azure Cosmos DB: Enables you to combine structured and unstructured data for analytics
  - Apache Kafka for HDInsight: Enables you to use Kafka as a streaming data source or sink
  - Azure Billing: You get a single bill from Azure
  - Azure Power BI: For rich data visualization
- Eliminates need to create a separate account with Databricks.

2. Data Acquisition and Understanding

Tools to understand data

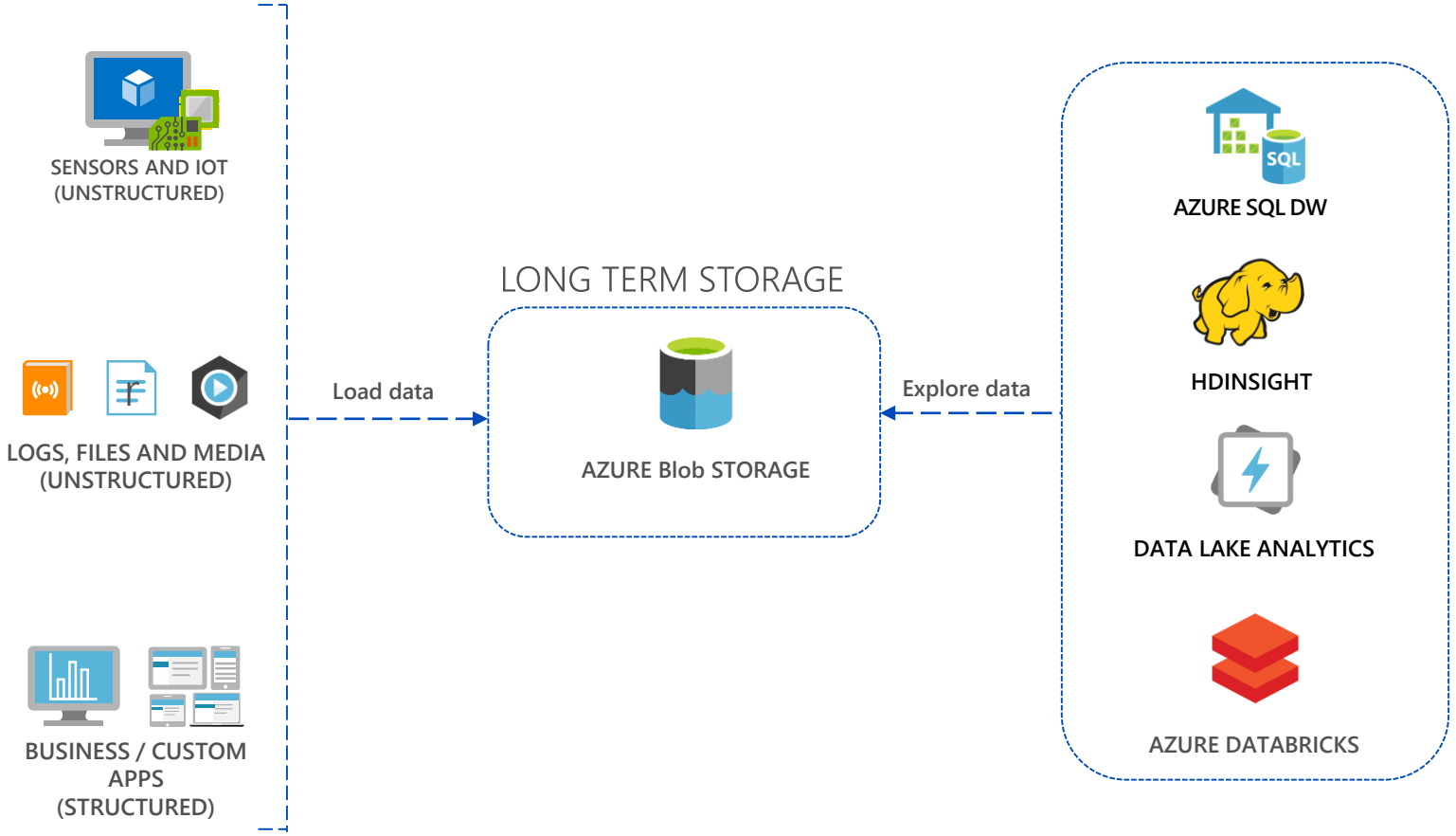
- Azure Databricks
  - Notebooks are not only for authoring Spark applications but can be *run/executed directly* on clusters
    - Shift+Enter
    - click the ▶ at the top right of the cell in a notebook
    - Submit via Job
  - Notebooks support fine grained permissions—so they can be *securely shared* with colleagues for collaboration (see following slide for details on permissions and abilities)
  - Notebooks are well-suited for prototyping, rapid development, exploration, discovery and iterative development



2. Data Acquisition and Understanding

Understand data

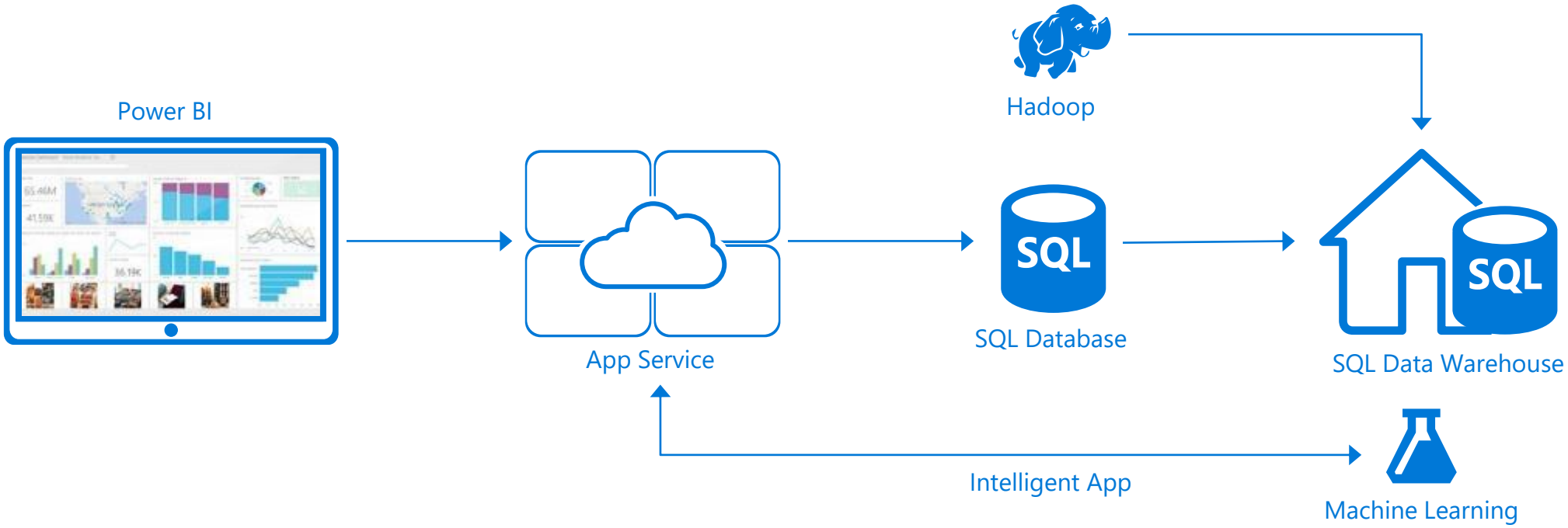
Prepare data



2. Data Acquisition and Understanding

Prepare data

- SQL Data Warehouse
  - Elastic data warehouse as a service with enterprise-class features



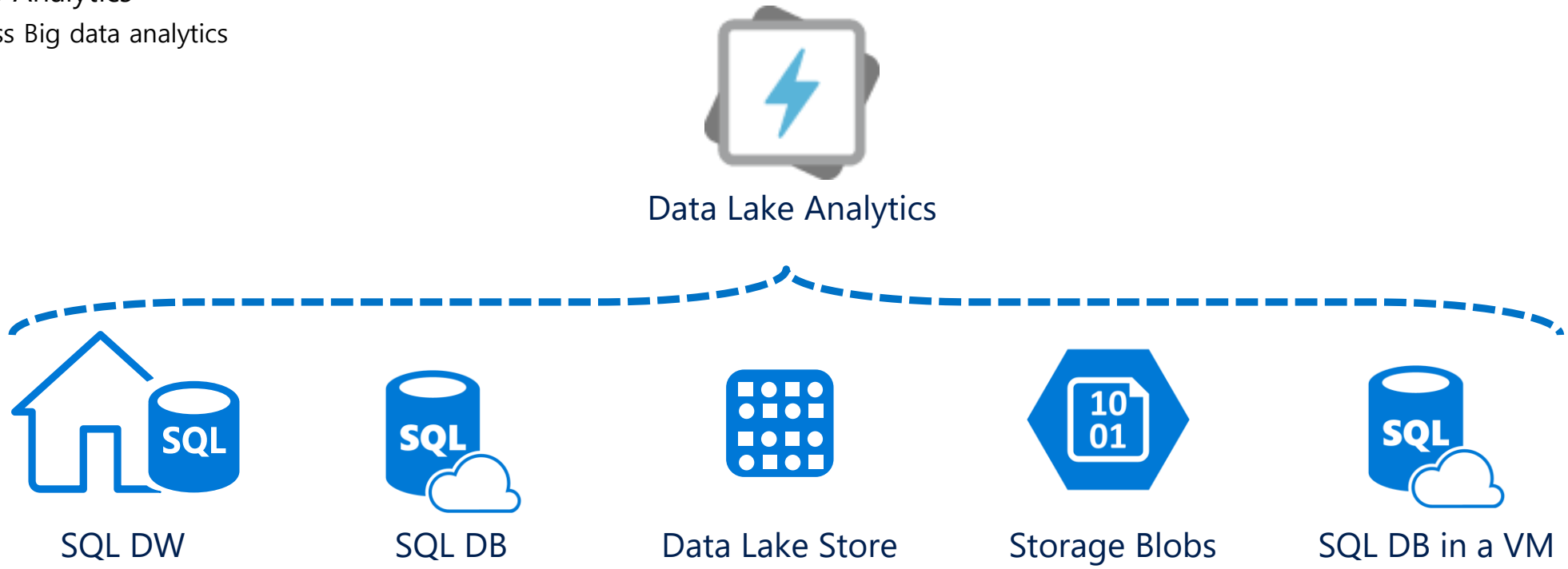
- Petabyte scale with massively parallel processing
- Independent scaling of compute and storage—in seconds
- Transact-SQL queries across relational and non-relational data

- Full enterprise-class SQL Server experience
- Works seamlessly with Power BI, Machine Learning, HDInsight, and Data Factory

2. Data Acquisition and Understanding

Prepare data

- Data Lake Analytics
  - Serverless Big data analytics



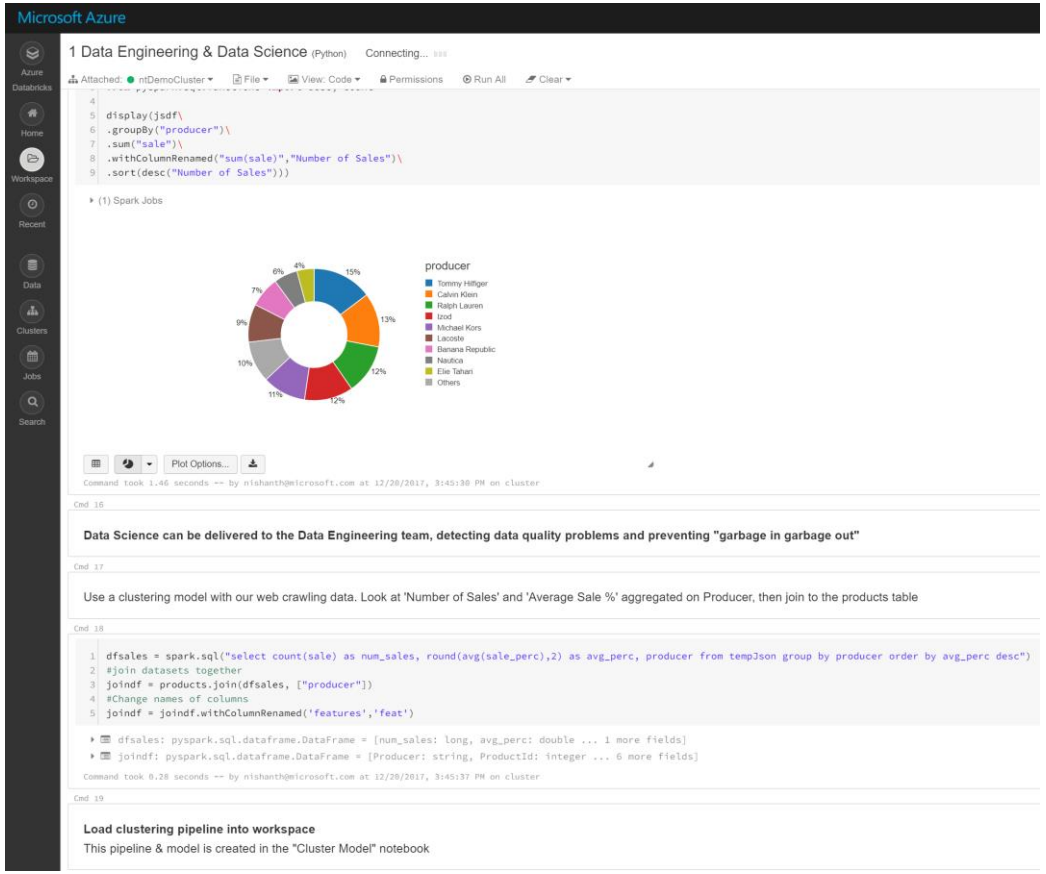
- Analyze data of any kind and size
- Develop faster, debug and optimize smarter
- Interactively explore patterns in your data
- No learning curve—use U-SQL, Spark, Hive, HBase and Storm

- Managed and supported with an enterprise-grade SLA
- Dynamically scales to match your business priorities
- Enterprise-grade security with Azure Active Directory
- Built on YARN, designed for the cloud

2. Data Acquisition and Understanding

Prepare data

- Azure Databricks
  - Fast and scalable data preparation for big data
  - Use a premium notebook environment
  - Leverage the full power of Spark to clean, curate and process data
  - Join with data from Cosmos DB and SQL DW, to manage master data and apply filt ers
  - Apply all transforms to massive amounts of data at scale within the same environme nt
  - Schedule this cleansing as a job



Data Science Tools

TDSP

Business Understanding	•Define Objectives •Identify Data Sources
Data Acquisition and Understanding	•Ingest Data •Explore Data •Update Data
Modeling	•Feature Selection •Create and Train Model
Deployment	•Operationalize
Customer Acceptance	•Testing and Validation •Handoff •Re-train and re-score

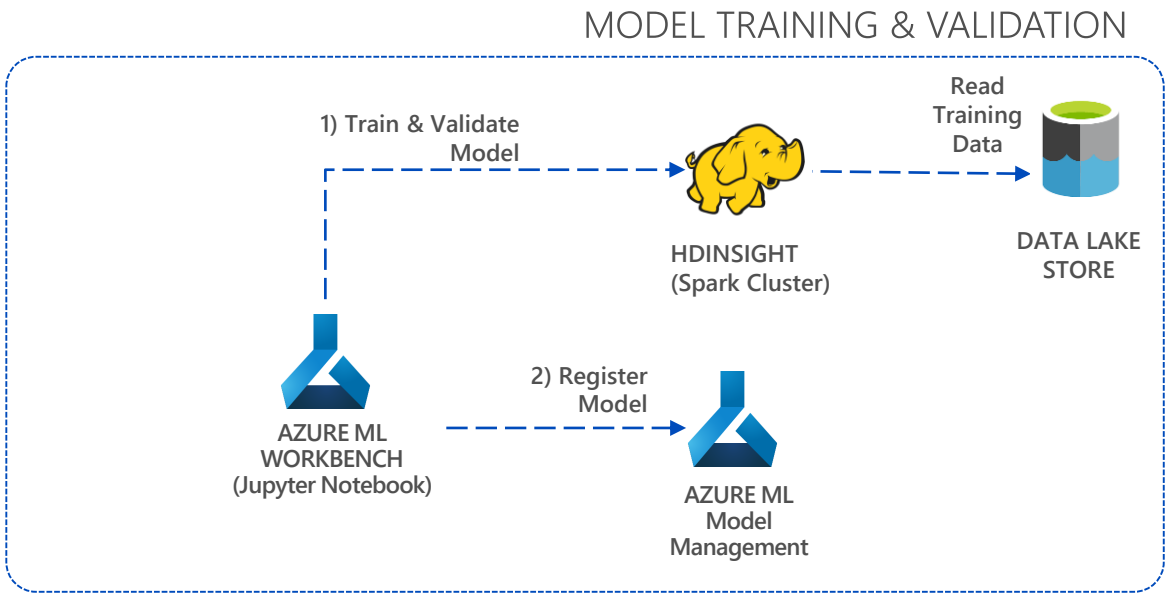
Azure Services

Information Management	• **Data Catalog** • **Data Factory** • **Event Hubs**
Big Data	• **Azure Storage** • **Data Lake** • **SQL Data Warehouse** • **Cosmos DB**
Intelligence and Advanced Analytics	• **Cortana, Bot Service, Cognitive Framework** • **Machine Learning** • **HDInsight** • **Data Lake** • **Stream Analytics** • **Analysis Services**
Visualization	• **Power BI** • **R**
Solutions	• **Templates and Gallery**

3. Modeling

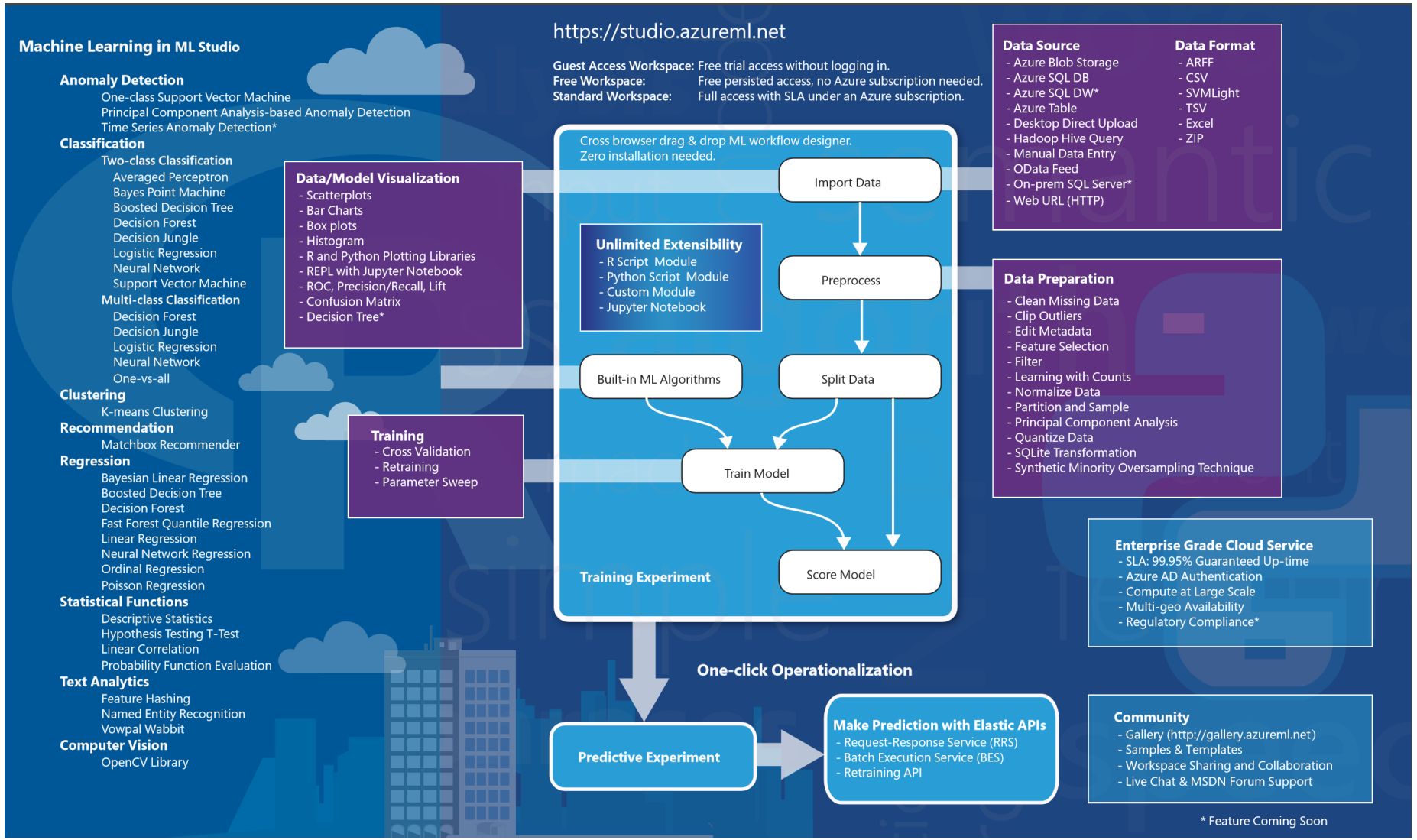
Model Selection and Training

Model Selection



3. Modeling

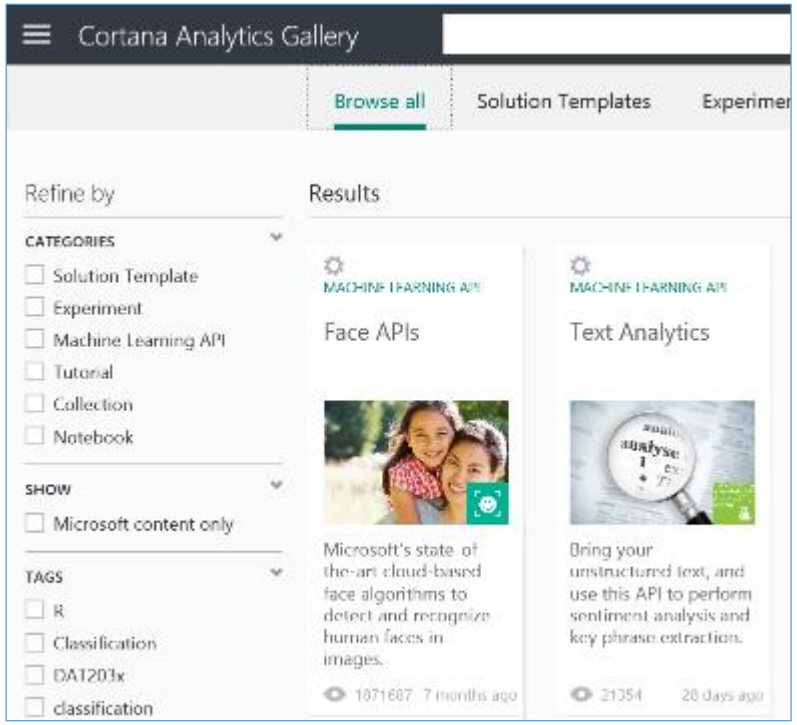
Model Selection and Training

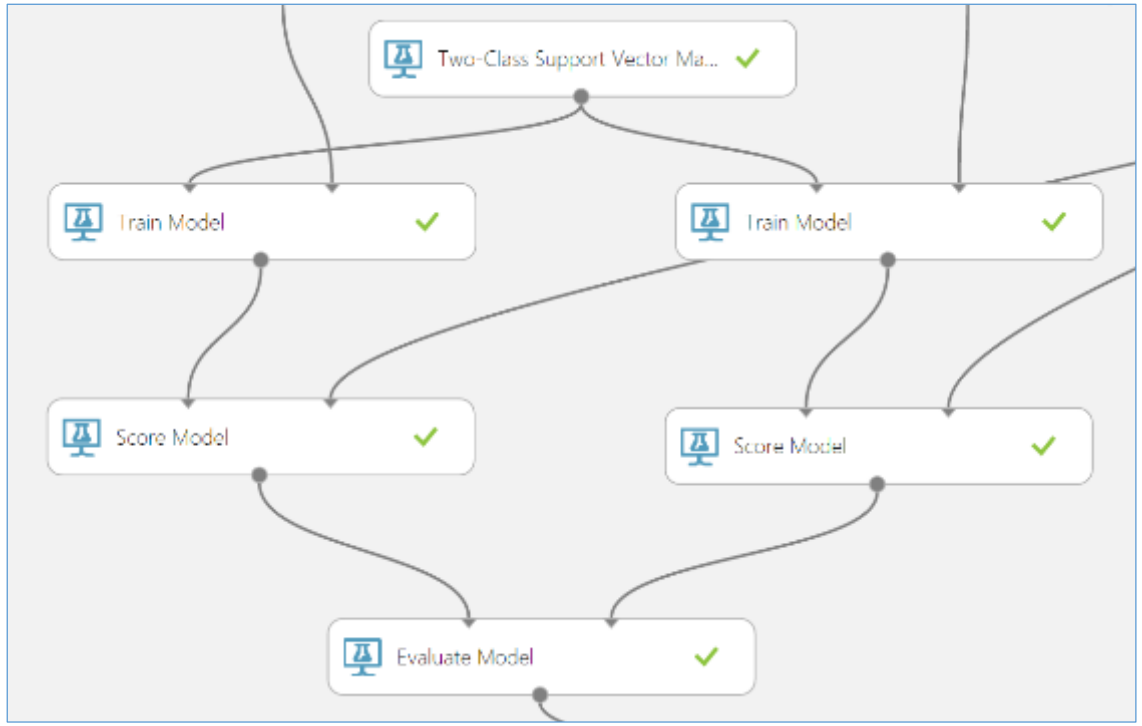


3. Modeling

Model Selection and Training

- Azure Machine Learning Studio



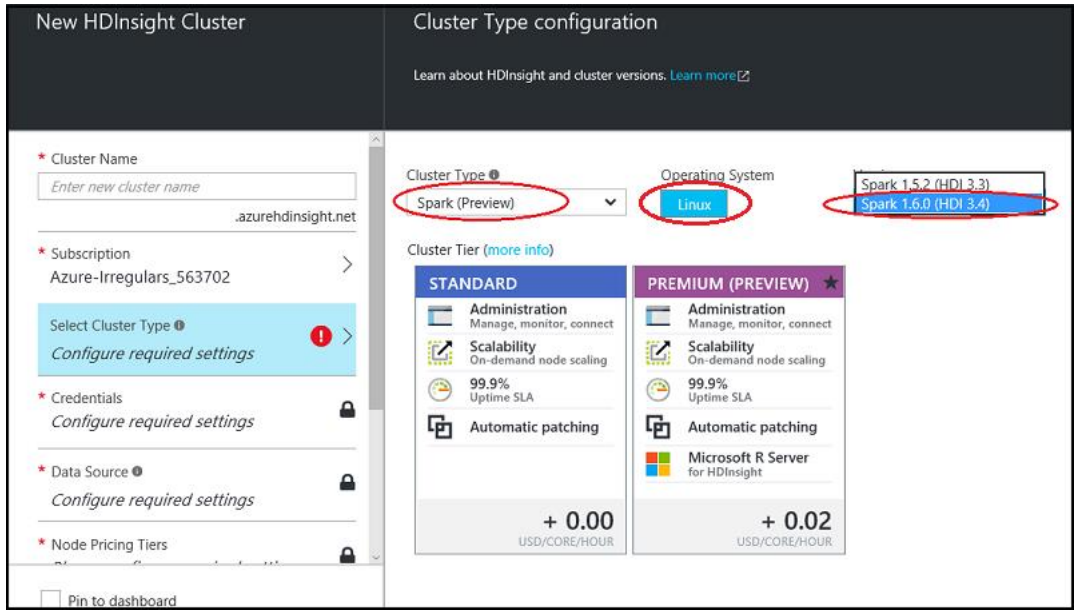


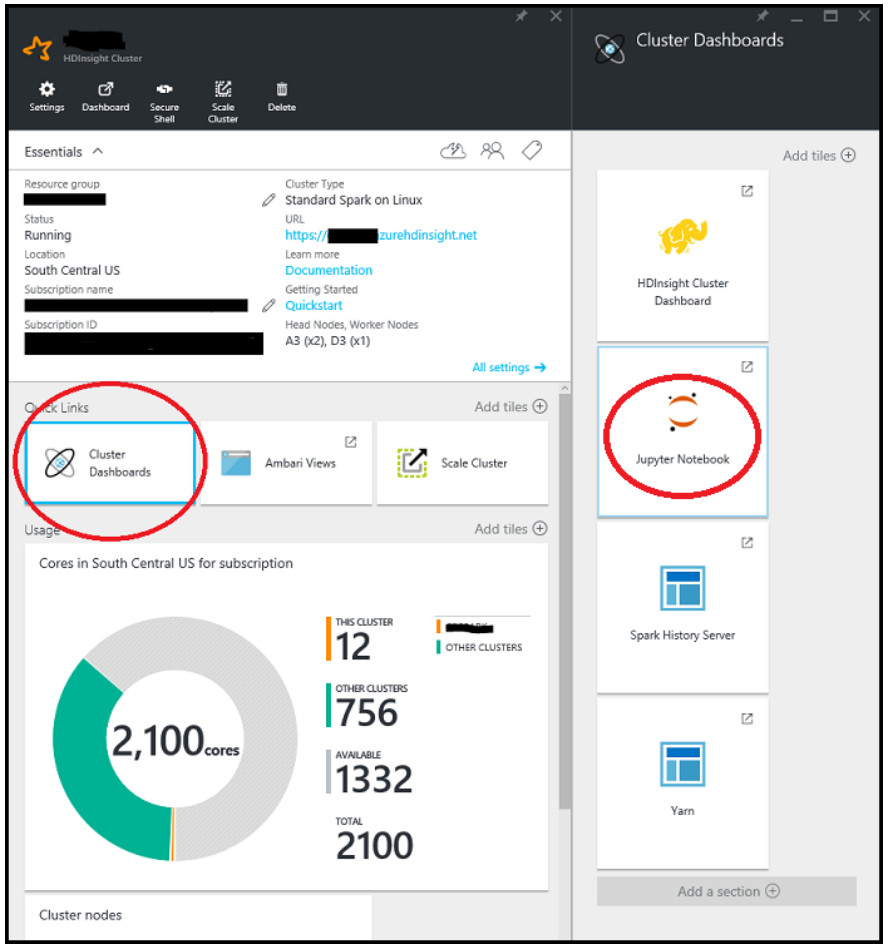
- Simple, scalable, cutting edge. A fully managed cloud service that enables you to easily build, deploy, and share predictive analytics solutions.
- Deploy in minutes. Azure Machine Learning means business. You can deploy your model into production as a web service that can be called from any device, anywhere and that can use any data source.
- Publish, share, monetize. Share your solution with the world in the Gallery or on the Azure Marketplace.

3. Modeling

Model Selection and Training

- HDInsight Spark

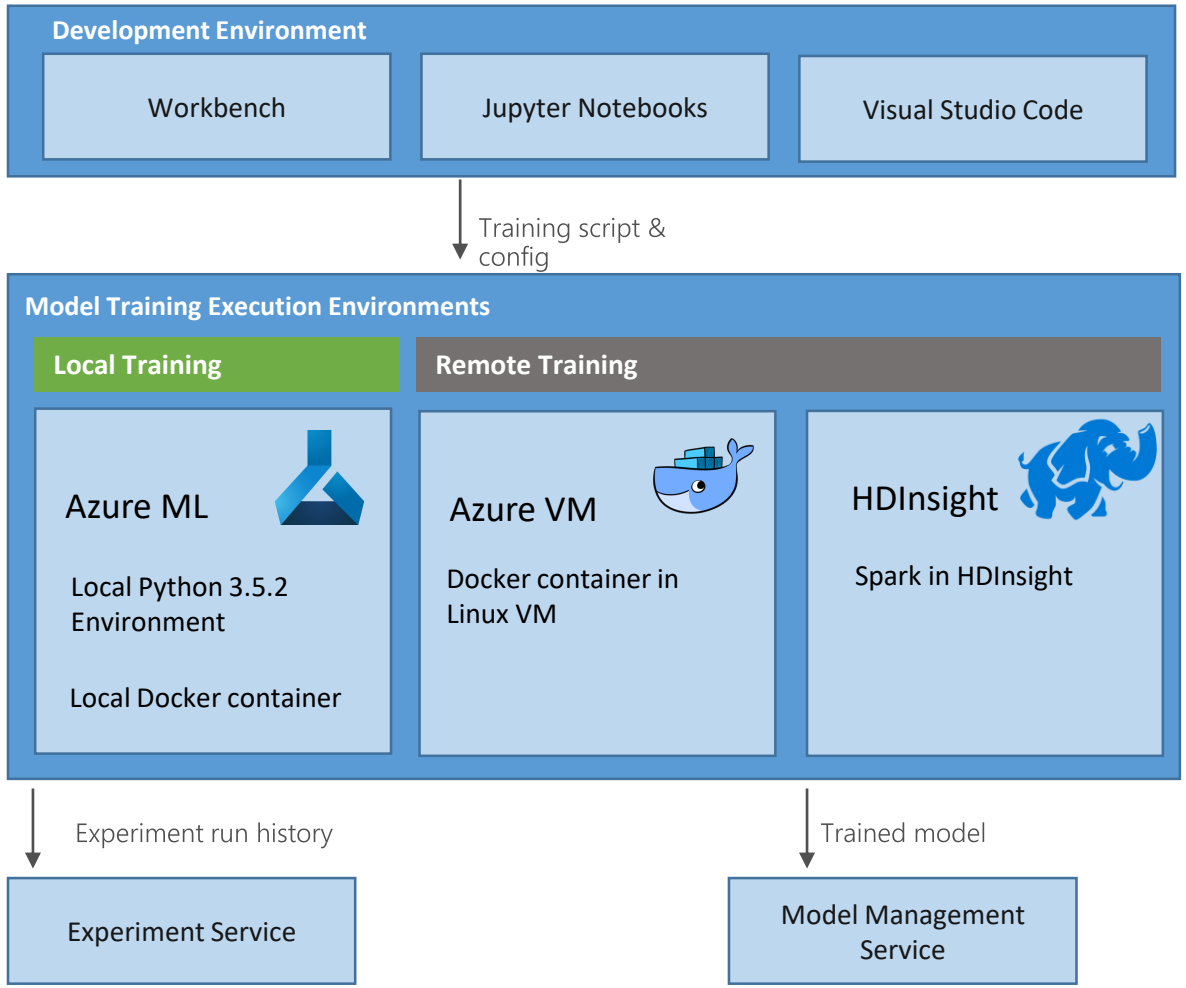




3. Modeling

Model Selection and Training

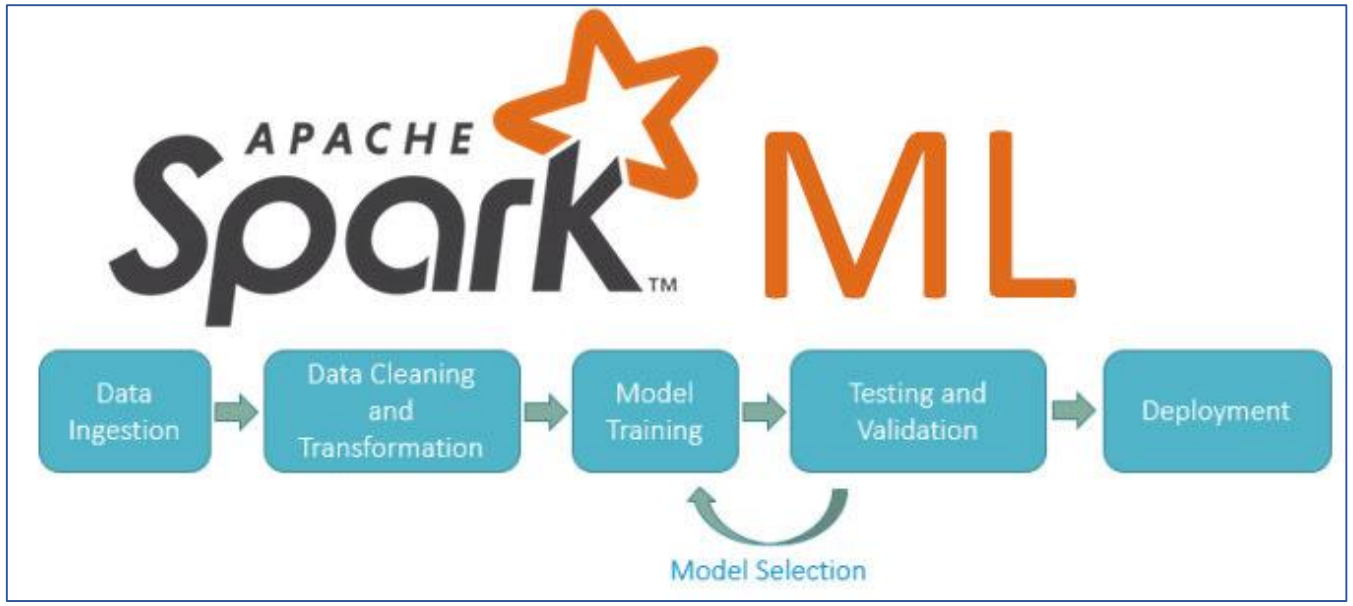
- Azure Machine Learning Workbench
  - Author Python training scripts using Jupyter Notebooks provide d within the Azure Machine Learning Workbench or with Visual Studio Code
  - Execute the training script on-premises or on a remote VM mach ine or HDInsight cluster
  - Experimentation Service handles execution of ML experiments ac ross environments, Git integration, access control, project roamin g and sharing and records run history information.
  - Model Management Service tracks model versions and lineage a cross training runs. Models are stored, registered, and managed i n the cloud.



3. Modeling

Model Selection and Training

- Azure Databricks
  - Offers a set of parallelized machine learning algorithms
  - Supports Model Selection (hyperparameter tuning) using Cross Validation and Train-Validation Split.
  - Supports Java, Scala or Python apps using DataFrame-based API (as of Spark 2.0). Benefits include:
    - An uniform API across ML algorithms and across multiple languages
    - Facilitates ML pipelines (enables combining multiple algorithms into a single pipeline).
    - Optimizations through Tungsten and Catalyst
  - Spark MLlib comes pre-installed on Azure Databricks and HDInsight
  - 3rd Party libraries supported include: H20 Sparkling Water, SciKit-learn and XGBoost



Data Science Tools

TDSP

Business Understanding	•Define Objectives •Identify Data Sources
Data Acquisition and Understanding	•Ingest Data •Explore Data •Update Data
Modeling	•Feature Selection •Create and Train Model
Deployment	•Operationalize
Customer Acceptance	•Testing and Validation •Handoff •Re-train and re-score

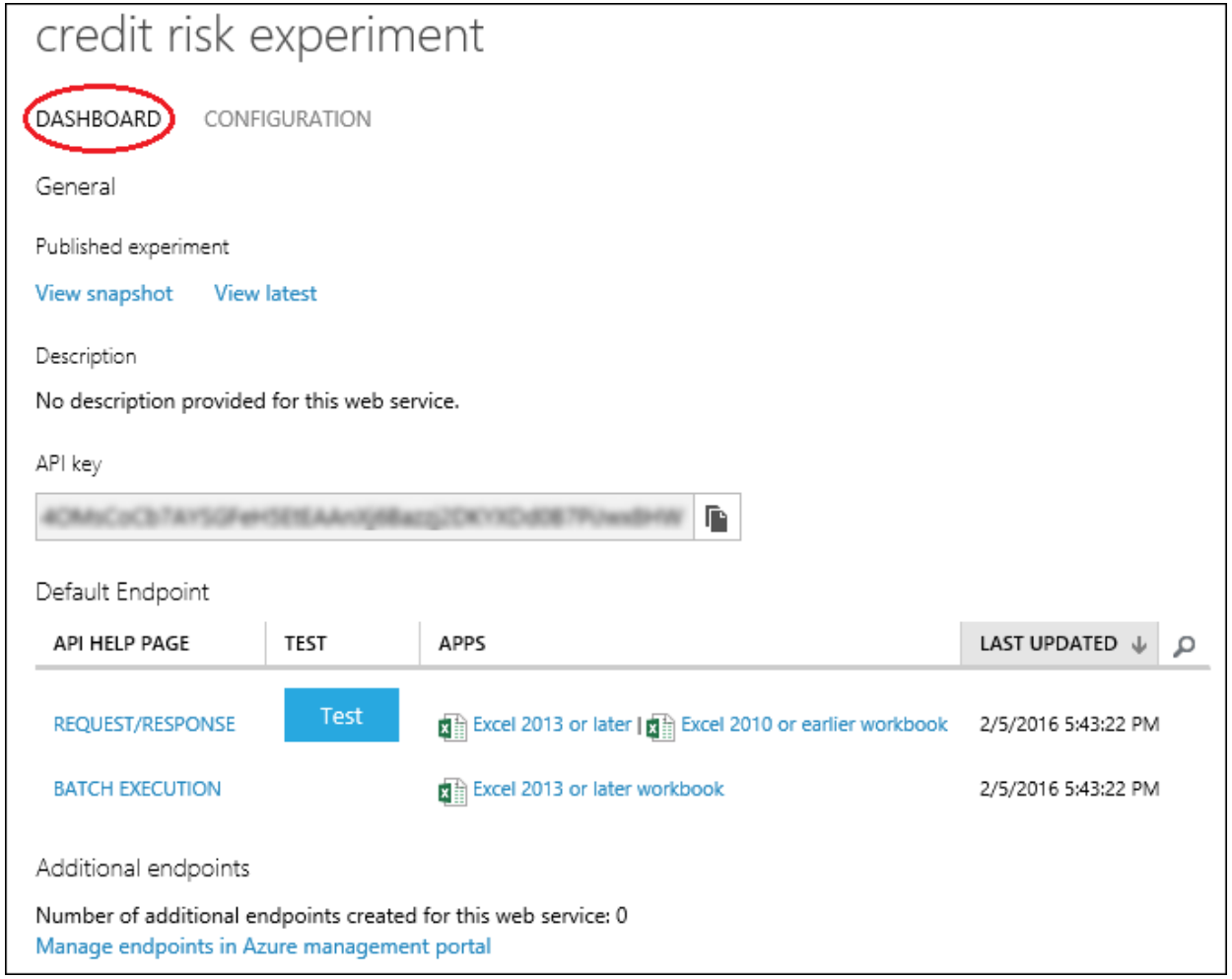
Azure Services

Information Management	• **Data Catalog** • **Data Factory** • **Event Hubs**
Big Data	• **Azure Storage** • **Data Lake** • **SQL Data Warehouse** • **Cosmos DB**
Intelligence and Advanced Analytics	• **Cortana, Bot Service, Cognitive Framework** • **Machine Learning** • **HDInsight** • **Data Lake** • **Stream Analytics** • **Analysis Services**
Visualization	• **Power BI** • **R**
Solutions	• **Templates and Gallery**

4. Deployment

Operationalize models

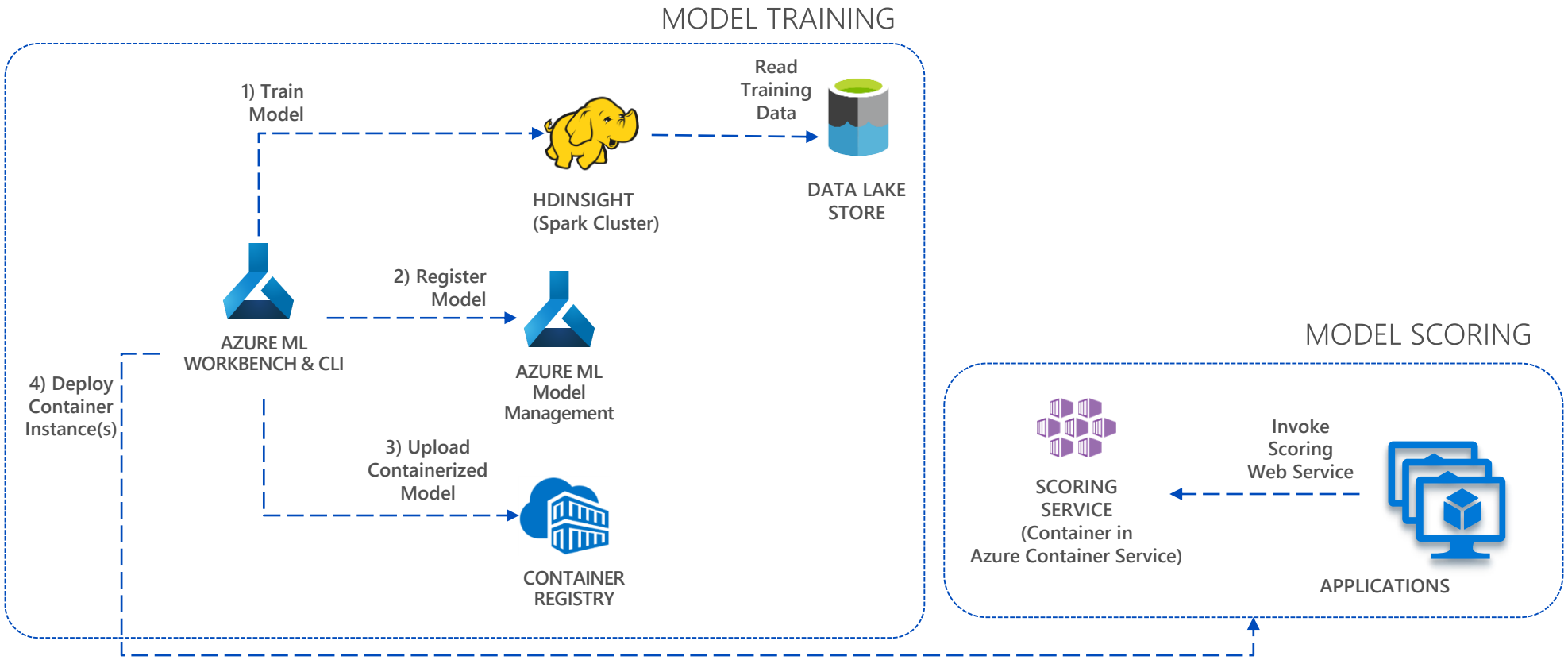
- Azure Machine Learning Studio - Publish as a Web Service



4. Deployment

Operationalize models

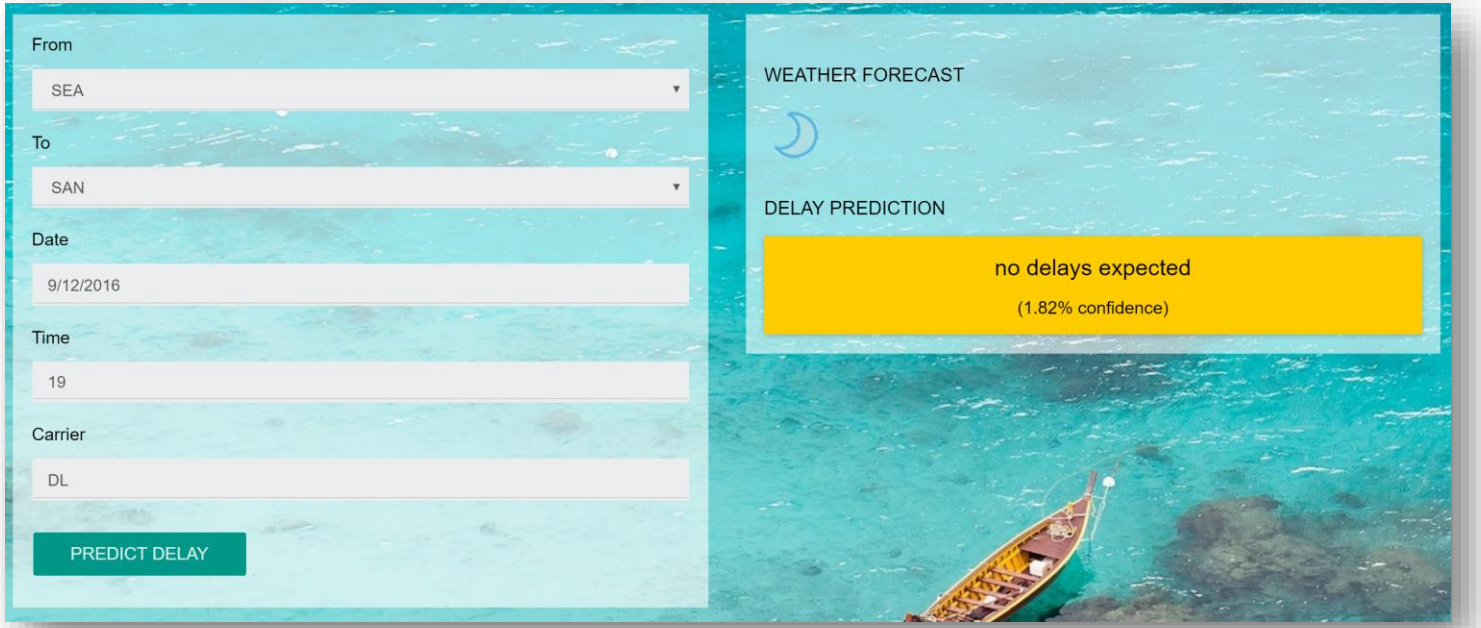
- Azure Machine Learning Services



4. Deployment

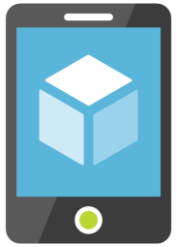
Operationalize models

- App Services



API App

Logic App

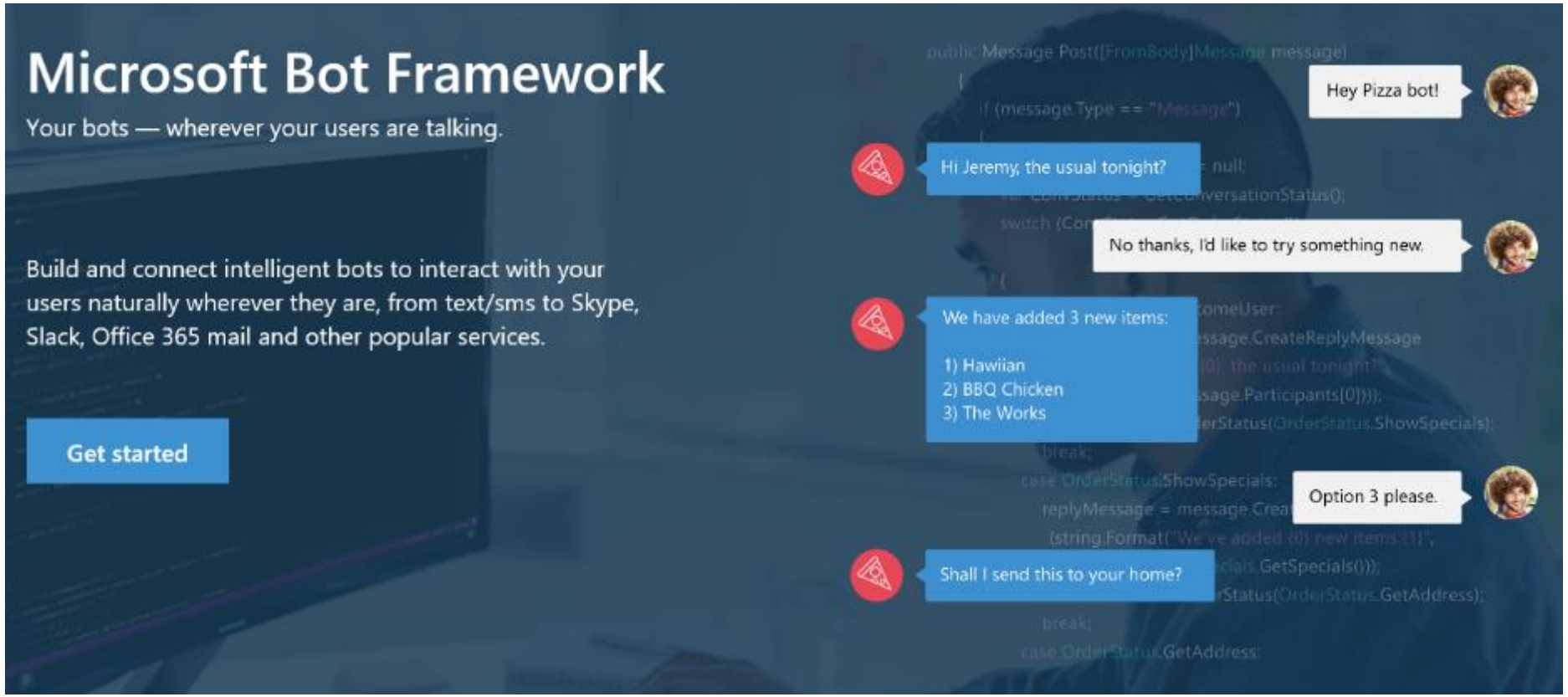


Mobile App

Web App

4. Deployment

Get things done in more helpful, proactive and natural ways



- Bot Connector Service: A service to register your bot, configure channels and publish to the Bot Directory. Connect your bot(s) seamlessly to text/sms, Office 365 mail, Skype, Slack, Twitter, and more.
- Bot Builder SDK: An open source SDK hosted on GitHub. Everything you need to build great dialogs within your Node.js or C# bot
- Bot Directory: A public directory of bots registered through the Bot Connector Service. Discover, try, and add bots to conversation experiences

4. Deployment

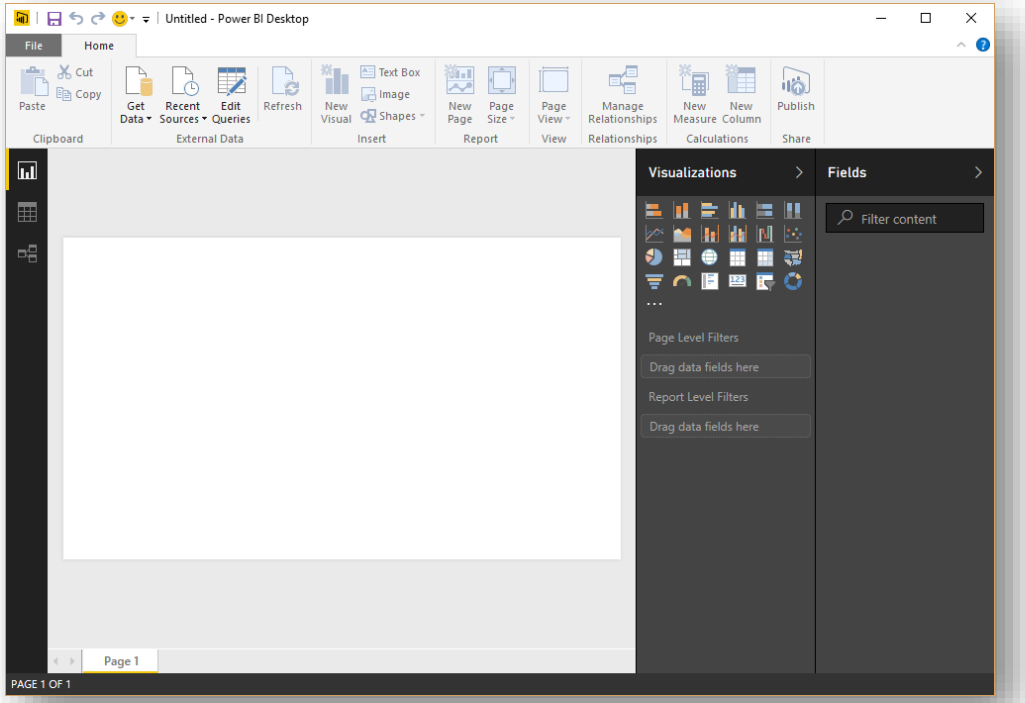
Get things done in more helpful, proactive and natural ways

Here are some of the things I can help you with…	Cortana for Consumers (today)	With the Cortana Intelligence Suite
Answers	Public reference data answers – *"How far is it from Los Angeles to San Francisco?"*	Answers from organizational data in Power BI *"What were our biggest deals that closed last month?"*
Predictions	Event predictions – *"Who do you think is going to win the Germany Italy game?"*	Integration with prediction solutions *"Which of our customers are most likely to churn in the next quarter?"*
Monitoring & Alerts	Flight status, traffic conditions, changes in weather, …	Monitoring KPIs and preemptive alerting *"Alert me if this customer ever has a 90% chance of churn in the next 30 days"*
Task Completion	Setting reminders, scheduling meetings, getting directions, …	Line of business process integration *Assistance with expense report submission on-time within policy*

4. Deployment

Report

- Power BI
  - A Reporting System for Multiple Data Sources
  - Available in:
    - Web Portal
    - Power BI Desktop
    - Microsoft Excel
    - Mobile apps (iOS, Android, Windows)
  - Author
    - Connect to Data
    - Shape the Data
    - Model the Data
    - Report on the Data
  - Publish
    - Local
    - To Service



Data Science Tools

TDSP

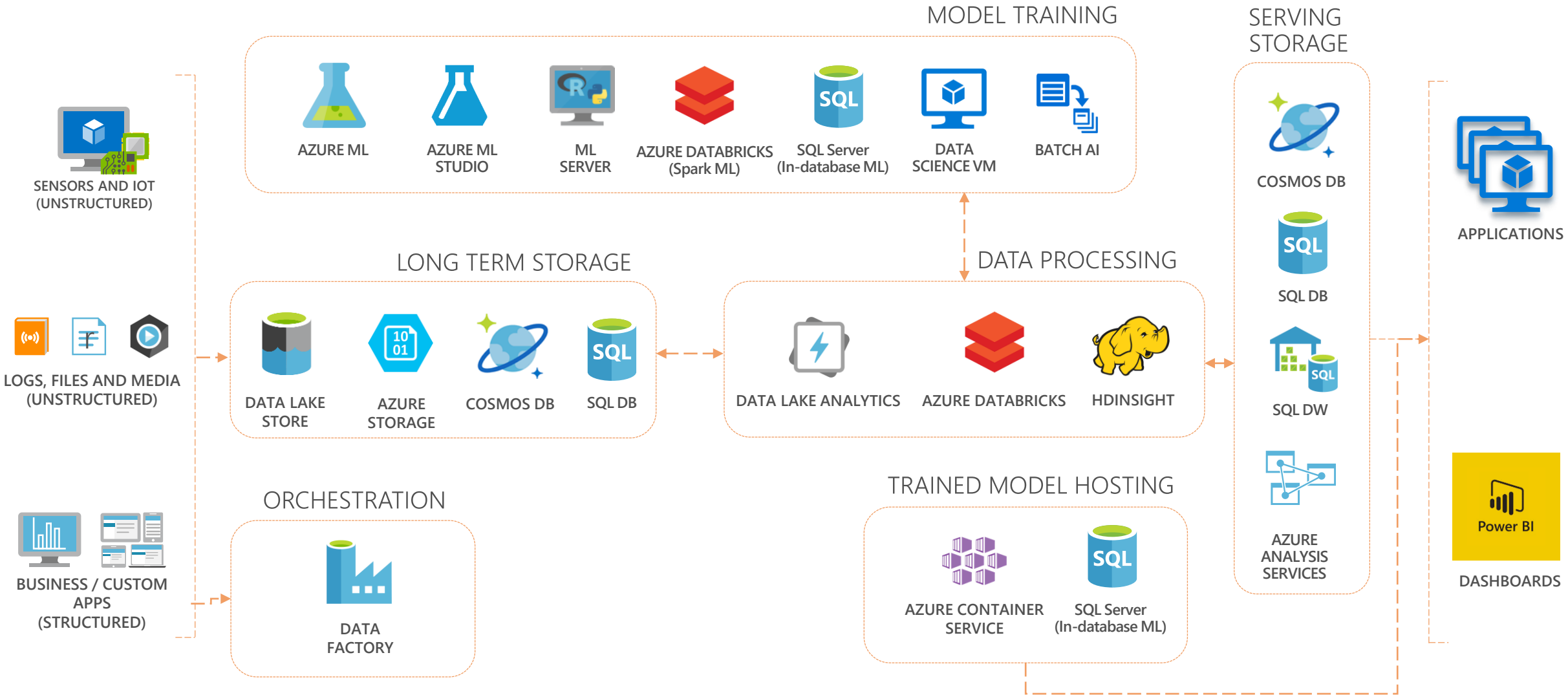
Business Understanding	•Define Objectives •Identify Data Sources
Data Acquisition and Understanding	•Ingest Data •Explore Data •Update Data
Modeling	•Feature Selection •Create and Train Model
Deployment	•Operationalize
Customer Acceptance	•Testing and Validation •Handoff •Re-train and re-score

Azure Services

Information Management	• **Data Catalog** • **Data Factory** • **Event Hubs**
Big Data	• **Azure Storage** • **Data Lake** • **SQL Data Warehouse** • **Cosmos DB**
Intelligence and Advanced Analytics	• **Cortana, Bot Service, Cognitive Framework** • **Machine Learning** • **HDInsight** • **Data Lake** • **Stream Analytics** • **Analysis Services**
Visualization	• **Power BI** • **R**
Solutions	• **Templates and Gallery**

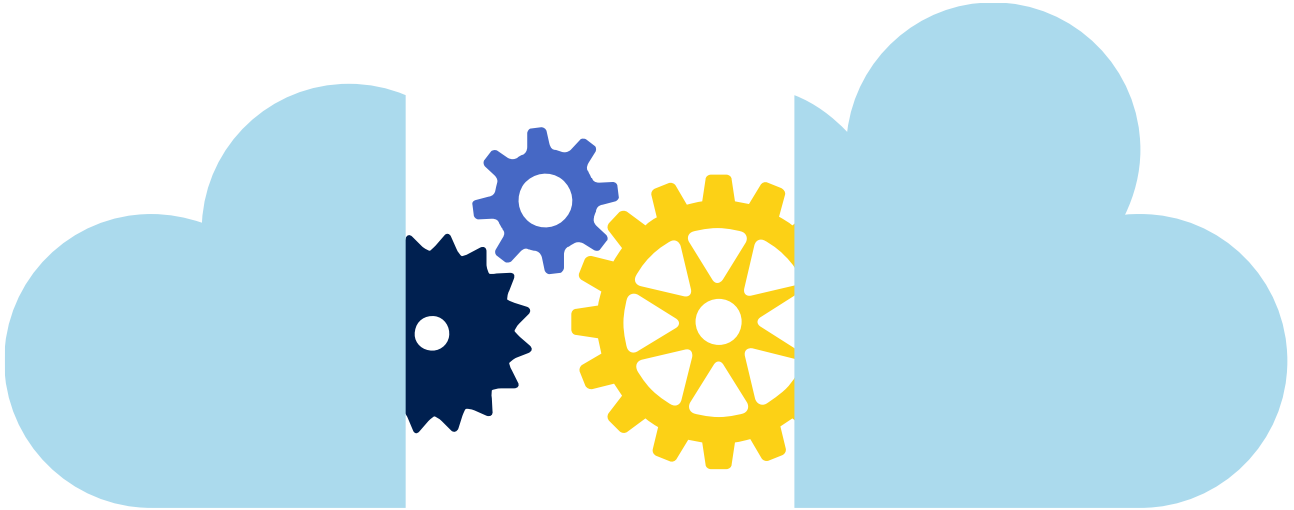
Advanced Analytics Pattern in Azure

Performing data collection/understanding, modeling and deployment



Agenda

- Data Science process – TDSP (Team Data Science Process)
- Data Science Tools
- **Hands on Lab**



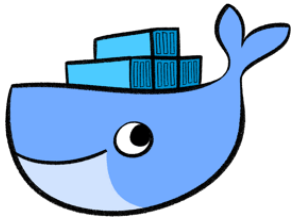
Hands on lab

Hands on lab

https://aka.ms/mtcs-azureml

1. Create DSVM for Windows Server

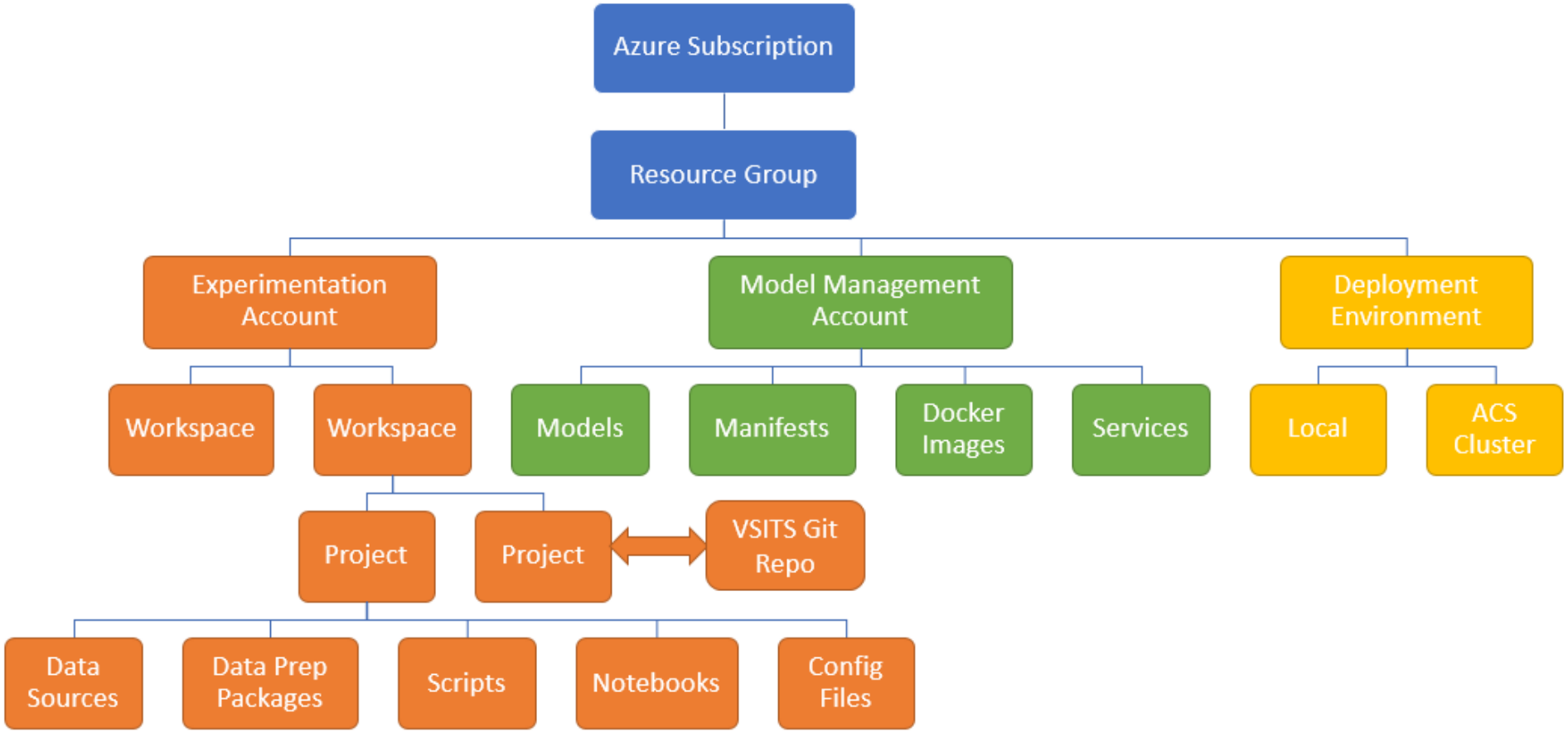
Azure Machine Learning Workbench



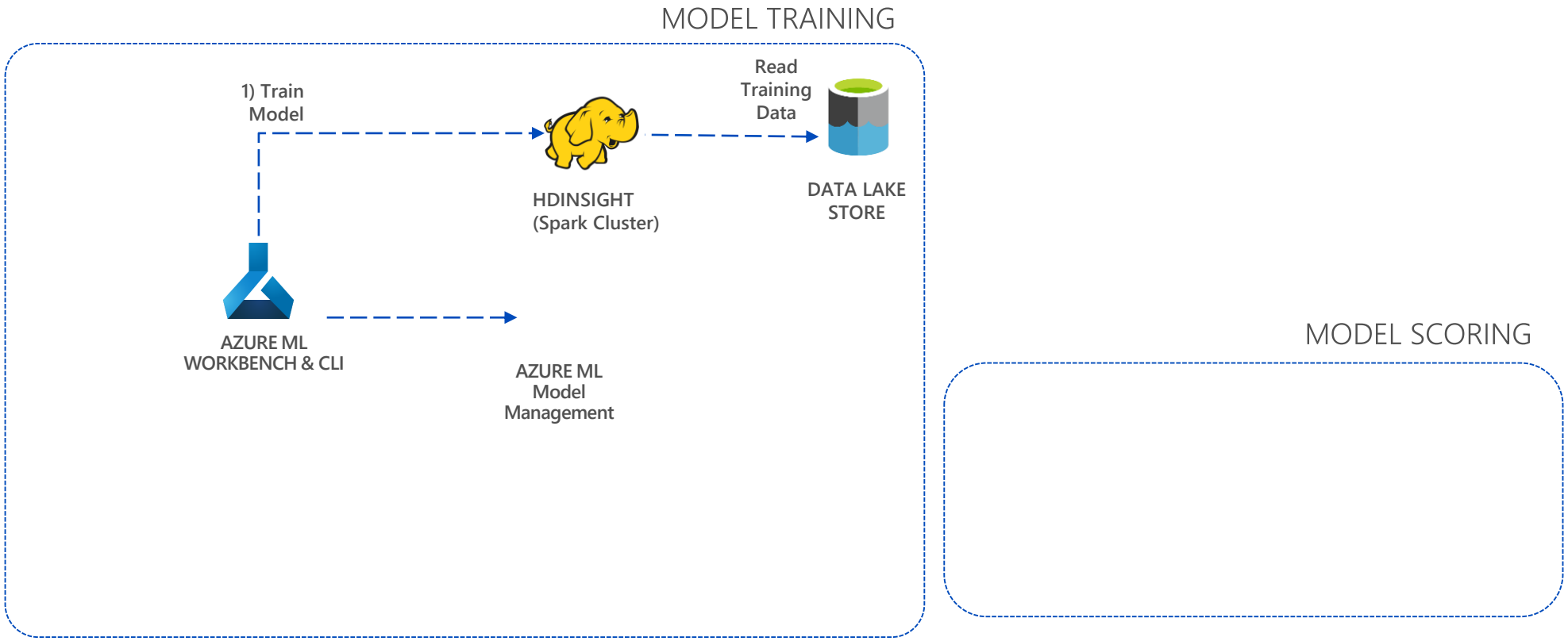
Docker for Windows

Data Science Virtual Machine for Windows 2016

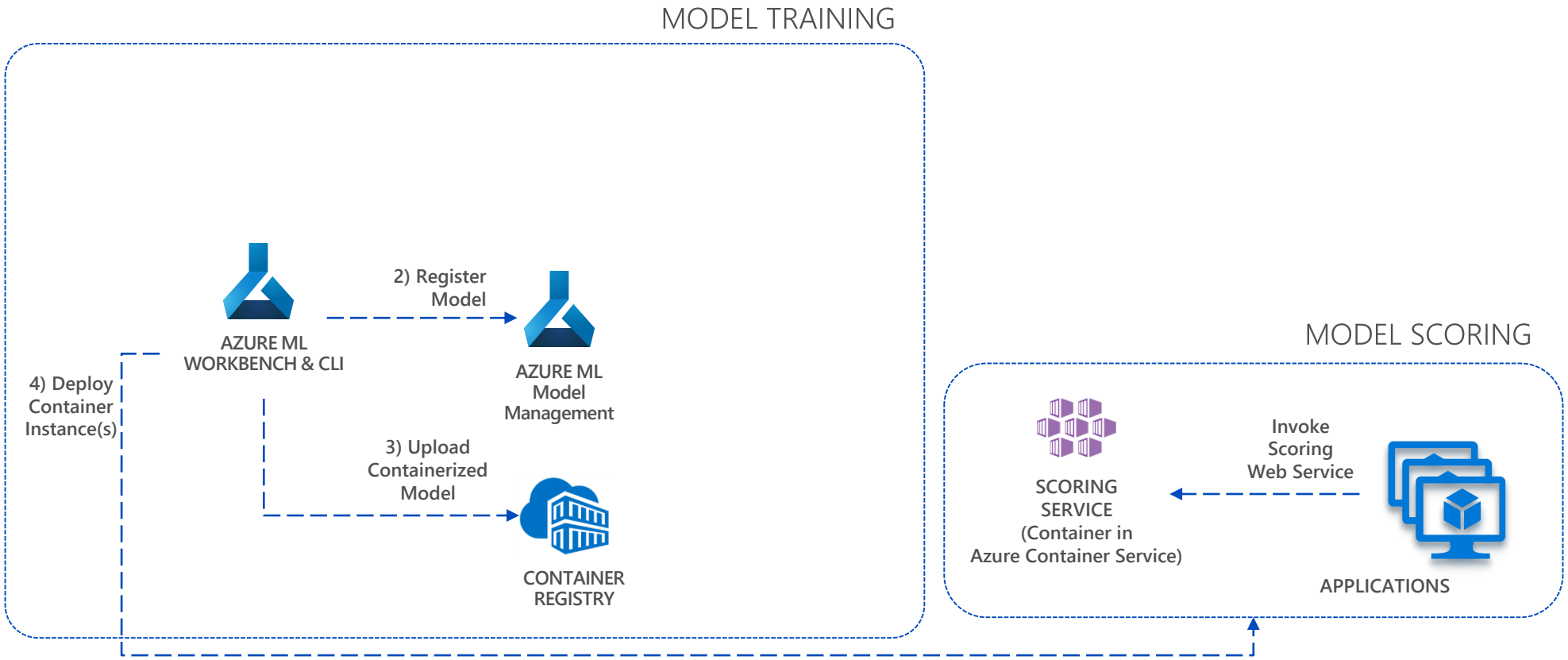
3. Install Azure Machine Learning Workbench



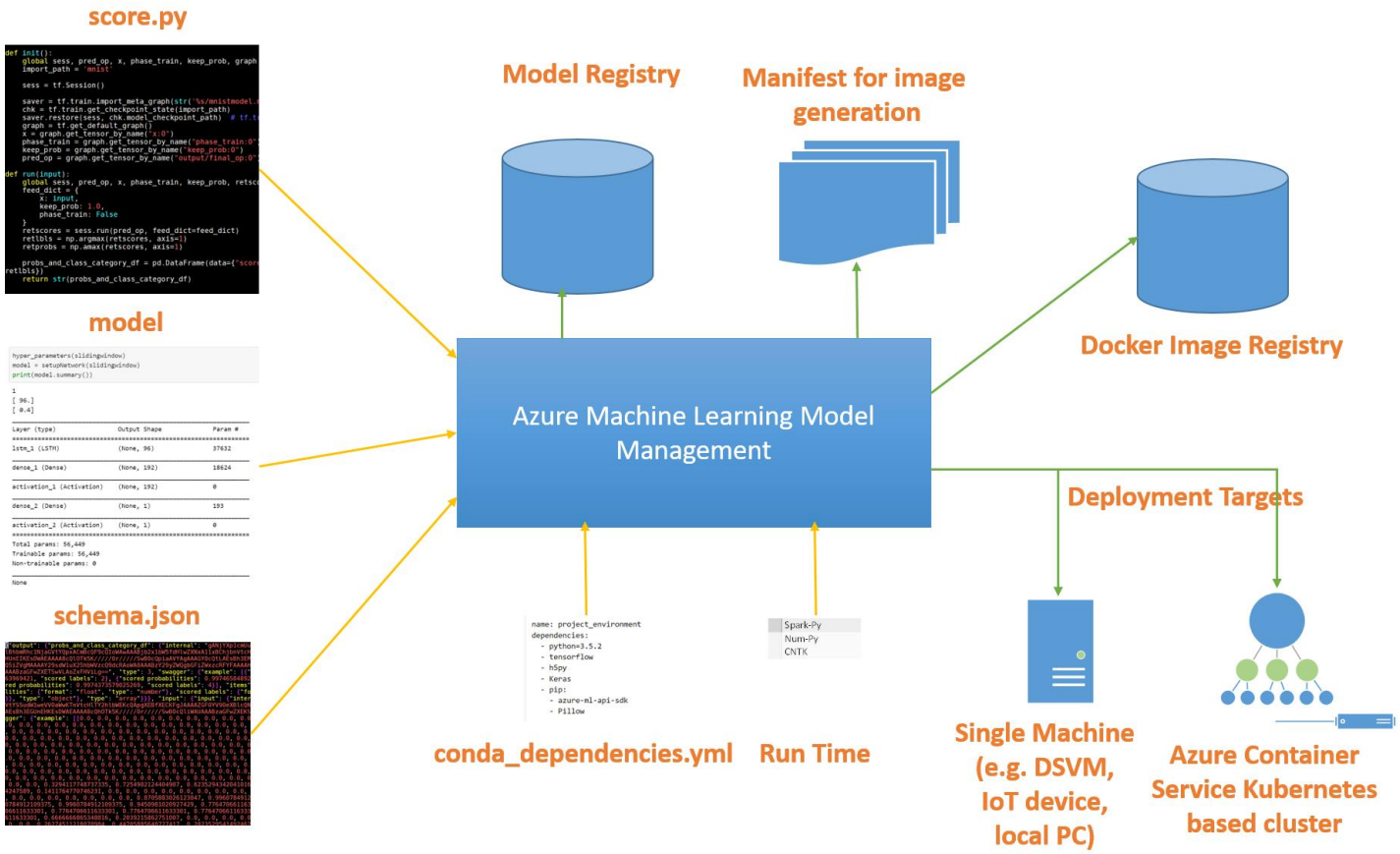
4. Model Selection



5. Deploy Model



7. Deploy Model



Microsoft

Microsoft Azure

Appendix

- Data Science for Beginners
  - https://docs.microsoft.com/en-us/azure/machine-learning/machine-learning-data-science-for-beginners-the-5-questions-data-science-answers

- R quick start
  - https://cran.r-project.org/doc/contrib/Paradis-rdebuts_en.pdf
  - https://docs.microsoft.com/en-us/azure/machine-learning/machine-learning-r-quickstart

- Setup Data Science Virtual Machine
  - https://docs.microsoft.com/en-us/azure/machine-learning/machine-learning-data-science-virtual-machine-overview