Microsoft Cloud Workshop

Big data and visualization

Julius Chen

Abstract and learning objectives

Abstract

Deploy a web app using Machine Learning to predict travel delays given flight delay data and weather conditions. Plan a bulk data import operation, followed by preparation, such as cleaning and manipulating the data for testing, and training your Machine Learning model.

Learning objectives

- Build a complete Azure Machine Learning (ML) model.
- Integrate an Azure ML web service into a Web App.
- Use Azure Data Factory (ADF) for data movement and operationalizing ML scoring.
- Summarize data with HDInsight and Spark SQL.
- Visualize batch predictions on a map using Power BI.

Step 1: Review the customer case study

Outcome

Analyze your customer needs

Timeframe

15 minutes

Customer situation

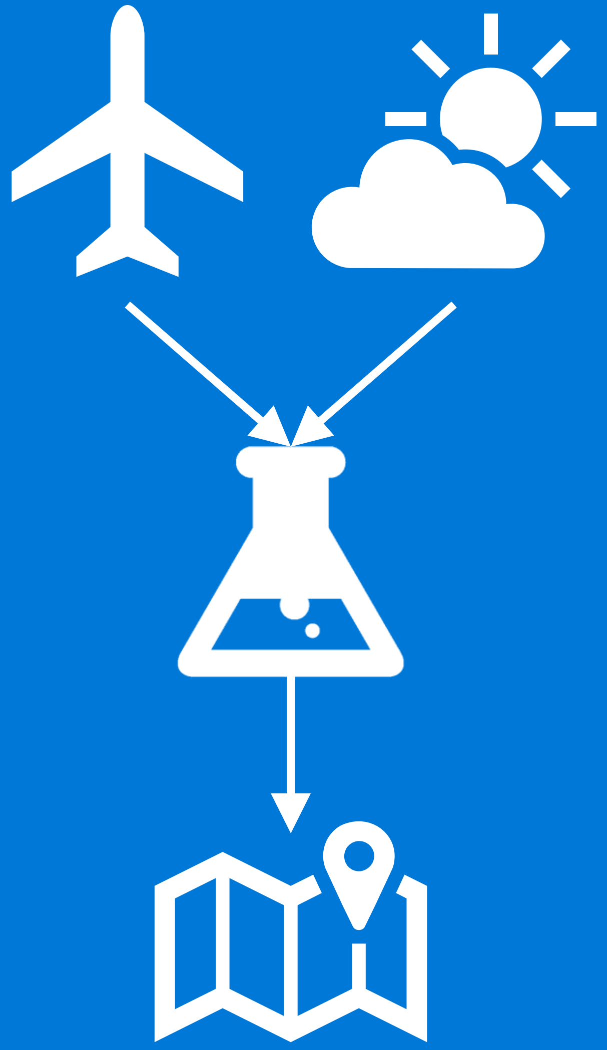
ADVENTUREWORKS TRAVEL

AdventureWorks Travel (AWT) provides concierge services for business travelers.

Interested in using predictive analytics to differentiate themselves in an increasingly crowded market.

Customer situation

- Proposed solution to provide flight delay risk assessment to customers
- Plan to use 30 years of flight delay and weather data
- Want to pilot the solution internally



Customer needs

- Modernize their analytics platform
- Ability to query data using SQL
- Load and store all data in Azure
- Use current weather forecast for flight delay predictions
- Proof of concept machine learning model
- Web-based visualizations of flight delay predictions

Customer objections

- Does Azure Machine Learning require a PhD in statistics?
- How long does it take to create and operationalize a machine learning model?
- Can operationalized ML models be flexible in the inputs they support?

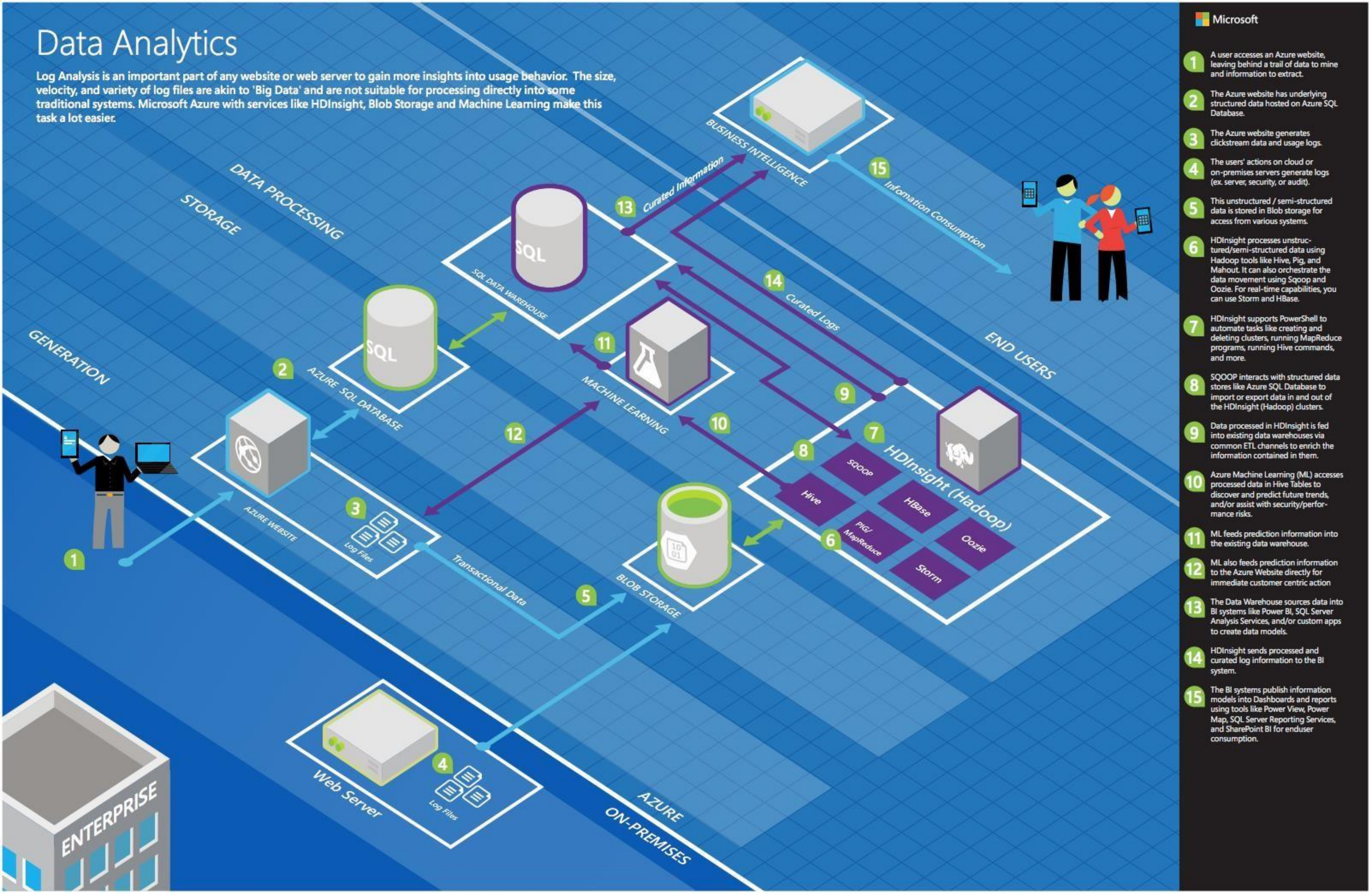
Customer objections

- What are the options for running SQL on Hadoop solutions in Azure?
- Does Azure offer anything to speed up querying files in HDFS?
- How can we identify, monitor, and protect PII data?

Customer objections

- Is Azure Data Lake a good fit for our PoC?
- Can access to our SQL DW be limited using Azure Active Directory?
- What data visualization tools are available on Azure? Can access to these be managed with Active Directory?

Common scenarios



Data Analytics

Log Analysis is an important part of any website or web server to gain more insights into usage behavior. The size, velocity, and variety of log files are akin to 'Big Data' and are not suitable for processing directly into some traditional systems. Microsoft Azure with services like HDInsight, Blob Storage and Machine Learning make this task a lot easier.

Microsoft

1. A user accesses an Azure website, leaving behind a trail of data to mine and information to extract.
2. The Azure website has underlying structured data hosted on Azure SQL Database.
3. The Azure website generates clickstream data and usage logs.
4. The users' actions on cloud or on-premises servers generate logs (ex. server, security, or audit).
5. This unstructured / semi-structured data is stored in Blob storage for access from various systems.
6. HDInsight processes unstructured/semi-structured data using Hadoop tools like Hive, Pig, and Mahout. It can also orchestrate the data movement using Sqoop and Oozie. For real-time capabilities, you can use Storm and HBase.
7. HDInsight supports PowerShell to automate tasks like creating and deleting clusters, running MapReduce programs, running Hive commands, and more.
8. SQOOP interacts with structured data stores like Azure SQL Database to import or export data in and out of the HDInsight (Hadoop) clusters.
9. Data processed in HDInsight is fed into existing data warehouses via common ETL channels to enrich the information contained in them.
10. Azure Machine Learning (ML) accesses processed data in Hive Tables to discover and predict future trends, and/or assist with security/performance risks.
11. ML feeds prediction information into the existing data warehouse.
12. ML also feeds prediction information to the Azure Website directly for immediate customer centric action
13. The Data Warehouse sources data into BI systems like Power BI, SQL Server Analysis Services, and/or custom apps to create data models.
14. HDInsight sends processed and curated log information to the BI system.
15. The BI systems publish information models into Dashboards and reports using tools like Power View, Power Map, SQL Server Reporting Services, and SharePoint BI for enduser consumption.

Step 2: Design the solution

Outcome

Design a solution and prepare to present the solution to the target customer audience in a 10-minute chalk-talk format.

Timeframe

60 minutes

Business needs (10 minutes)	• Respond to questions outlined in your guide and list the answers on a flipchart.
Design (35 minutes)	• Design a solution for as many of the stated requirements as time allows. Show the solution on a flipchart.
Prepare (15 minutes)	• Identify any customer needs that are not addressed with the proposed solution. • Identify the benefits of your solution. • Determine how you will respond to the customer's objections. • Prepare for a 10-minute presentation to the customer.

Step 3: Present the solution

Outcome

Present a solution to the target customer in a 10-minute chalk-talk format

Timeframe

30 minutes (15 minutes for each team to present and receive feedback)

Directions

- Pair with another table
- One table is the Microsoft team and the other table is the customer
- The Microsoft team presents their proposed solution to the customer
- The customer asks one of the objections from the list of objections in the case study
- The Microsoft team responds to the objection
- The customer team gives feedback to the Microsoft team

Wrap-up

Outcome

Identify the preferred solution for the case study

Identify solutions designed by other teams

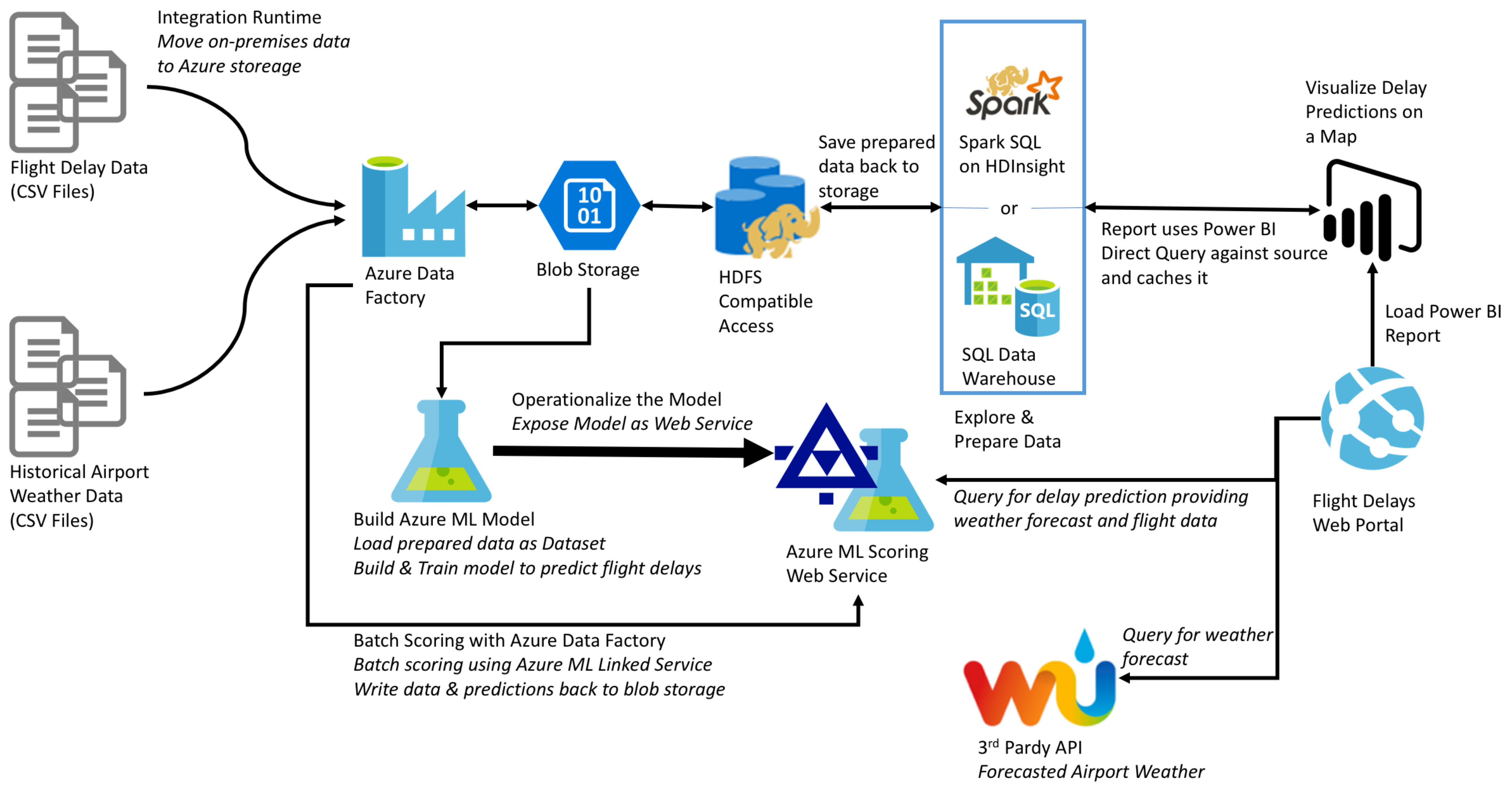
Timeframe

15 minutes

Preferred target audience

- Jack Tradewinds, CIO of AdventureWorks Travel
- The primary audience is business decision makers and technology decision makers.
- Usually we talk to the Infrastructure Managers who report to the CIOs, or to application sponsors (like a VP LOB, CMO) or to those that represent the Business Unit IT or developers that report to application sponsors.

Preferred solution



Integration Runtime
Move on-premises data to Azure storeage

Flight Delay Data (CSV Files)

Historical Airport Weather Data (CSV Files)

Azure Data Factory

Blob Storage

HDFS Compatible Access

Save prepared data back to storage

Spark SQL on HDInsight

or

SQL Data Warehouse

Explore & Prepare Data

Visualize Delay Predictions on a Map

Report uses Power BI Direct Query against source and caches it

Load Power BI Report

Flight Delays Web Portal

Operationalize the Model
Expose Model as Web Service

Build Azure ML Model
Load prepared data as Dataset
Build & Train model to predict flight delays

Azure ML Scoring Web Service

Query for delay prediction providing weather forecast and flight data

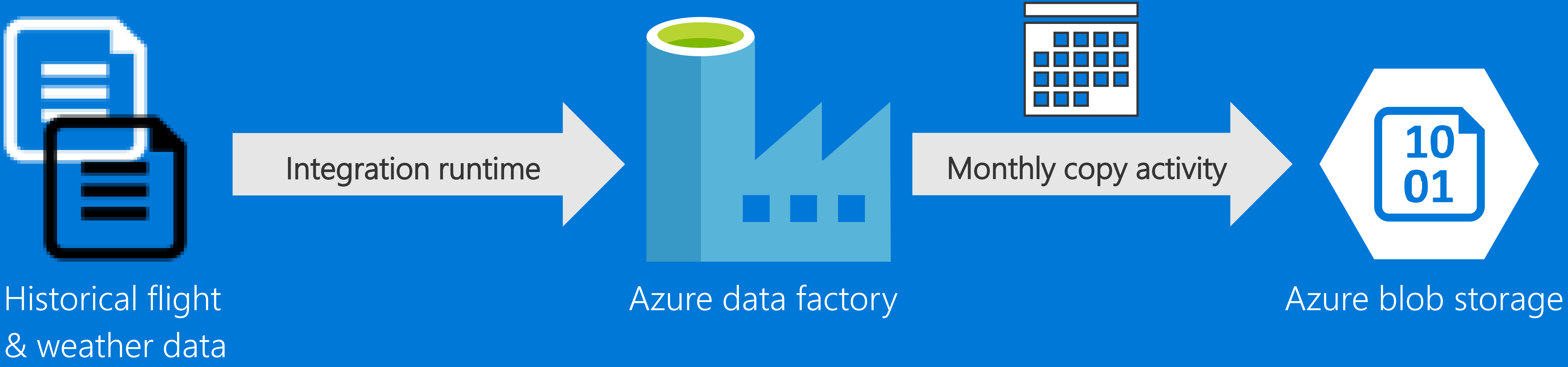
Query for weather forecast

Batch Scoring with Azure Data Factory
Batch scoring using Azure ML Linked Service
Write data & predictions back to blob storage

3rd Pardy API
Forecasted Airport Weather

Preferred solution

Data Loading



Historical flight & weather data

Integration runtime

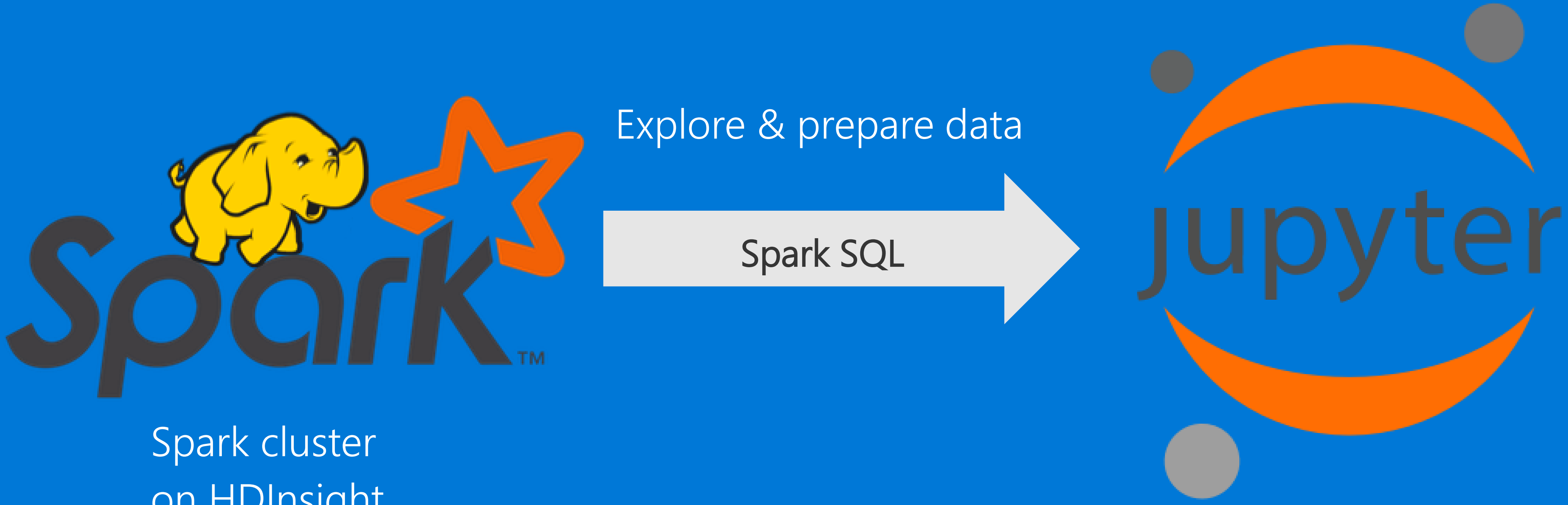
Azure data factory

Monthly copy activity

Azure blob storage

Preferred solution

Data reparation



Explore & prepare data

Spark SQL

Spark cluster
on HDInsight

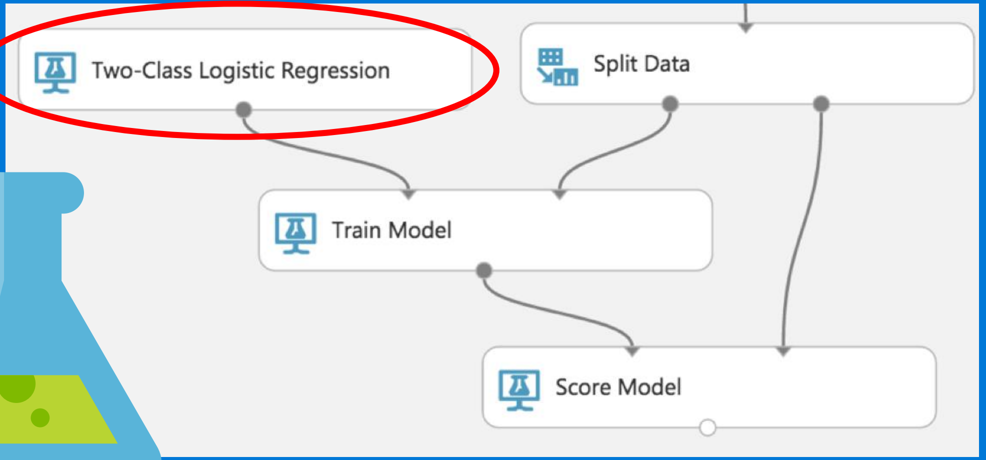
Jupyter notebook used
by AWT analysts

Preferred solution

Machine learning model

Prep data with Spark SQL

Spark cluster
on HDInsight

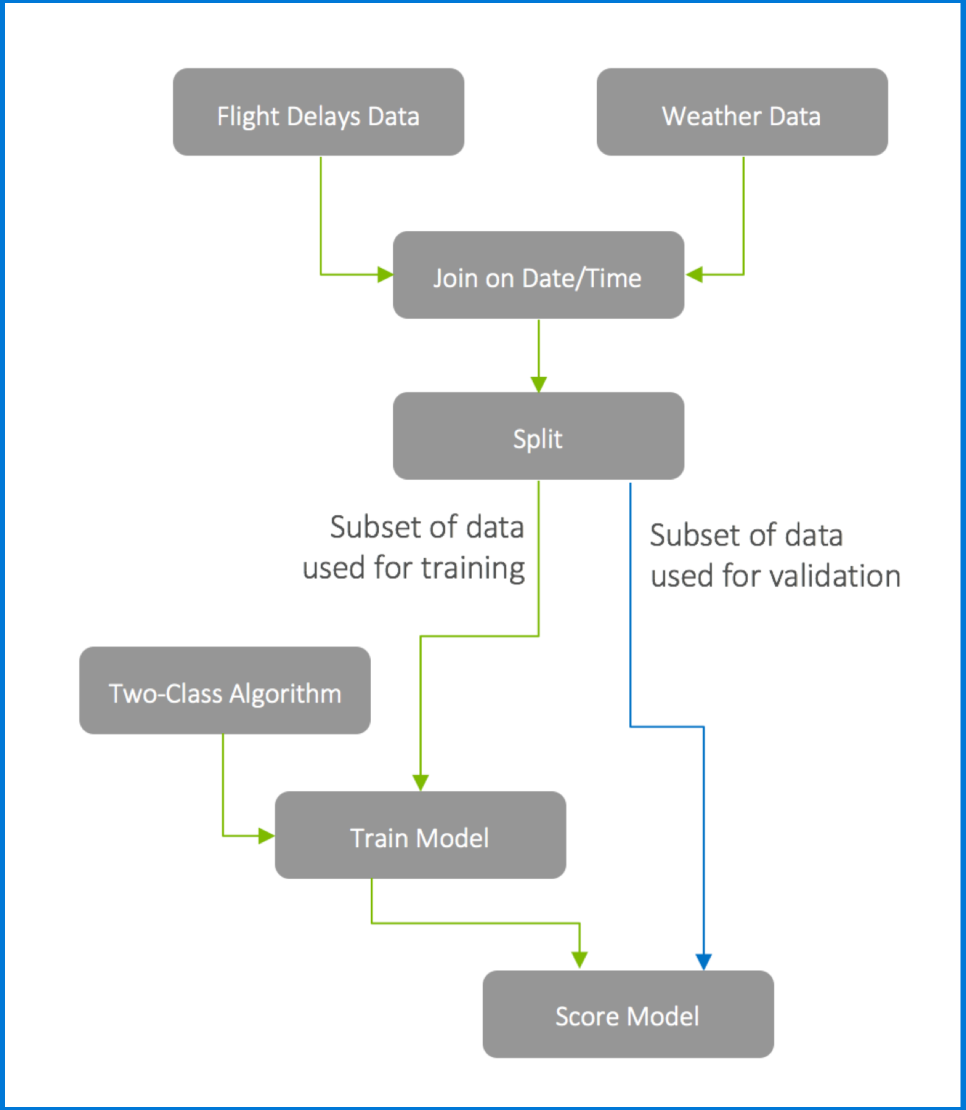


Azure Machine Learning (ML) with
Two-class logistic regression

Preferred solution

Machine learning model

- Start with domain knowledge
- Remove fields that do not add value
- Validate preliminary model against training data



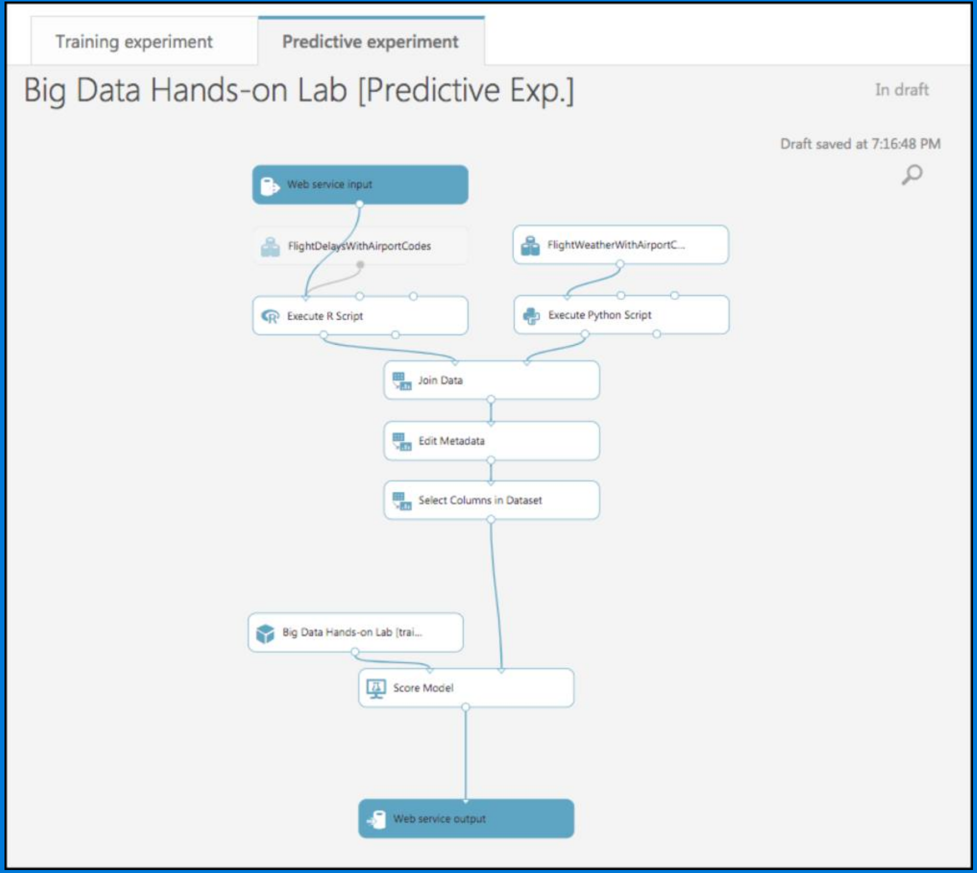
Preferred solution

Machine learning model

- Data lunging with R or Python
- Reserve some historical data to “test” the model
- Measure error on the training set and validation sets separately for indicator of whether model is in danger overfitting.

Preferred solution

Operationalizing machine learning



Training experiment | Predictive experiment

Big Data Hands-on Lab [Predictive Exp.]

In draft

Draft saved at 7:16:48 PM

Web service input
FlightDelaysWithAirportCodes
FlightWeatherWithAirportC...
Execute R Script
Execute Python Script
Join Data
Edit Metadata
Select Columns in Dataset
Big Data Hands-on Lab [trai...
Score Model
Web service output

Azure ML Studio

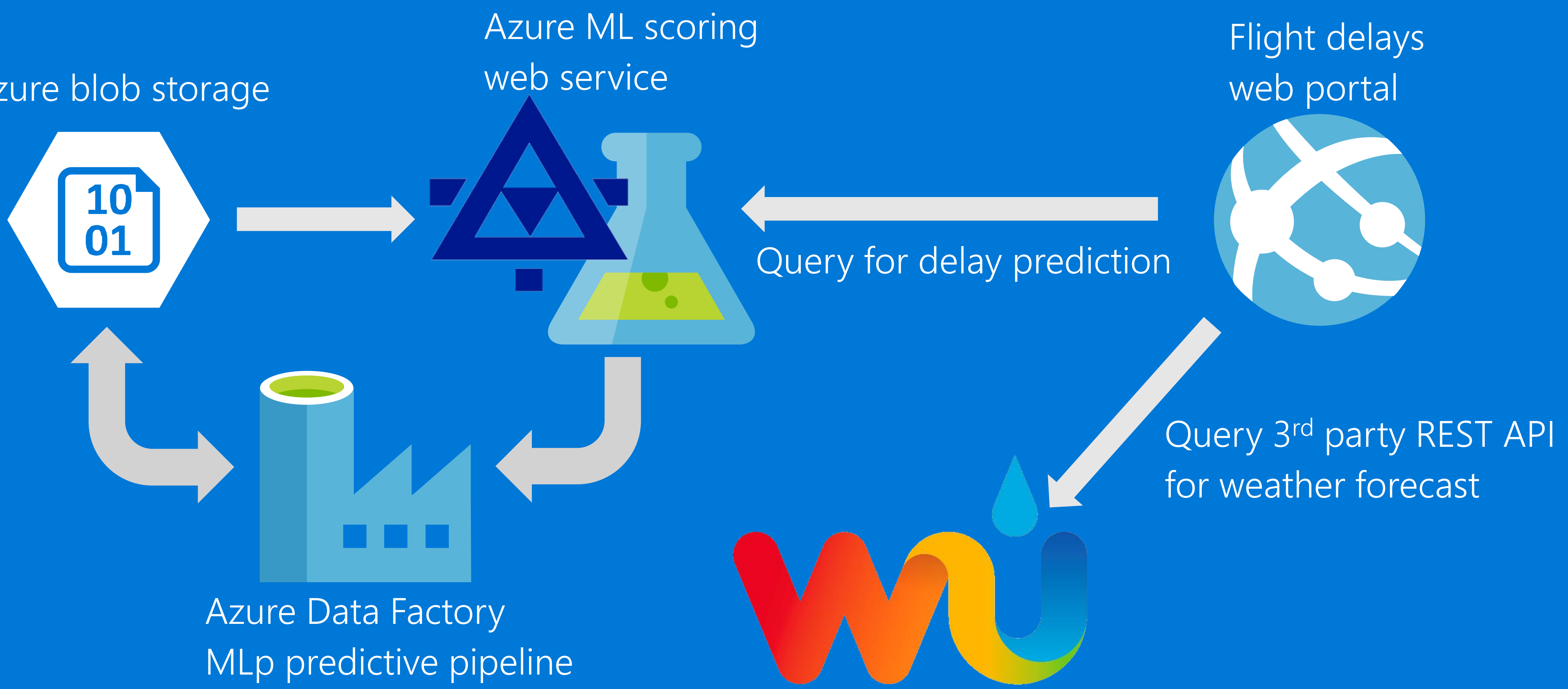
Publish via Azure ML Studio

Operationalize

Predictive web service
(REST API)

Preferred solution

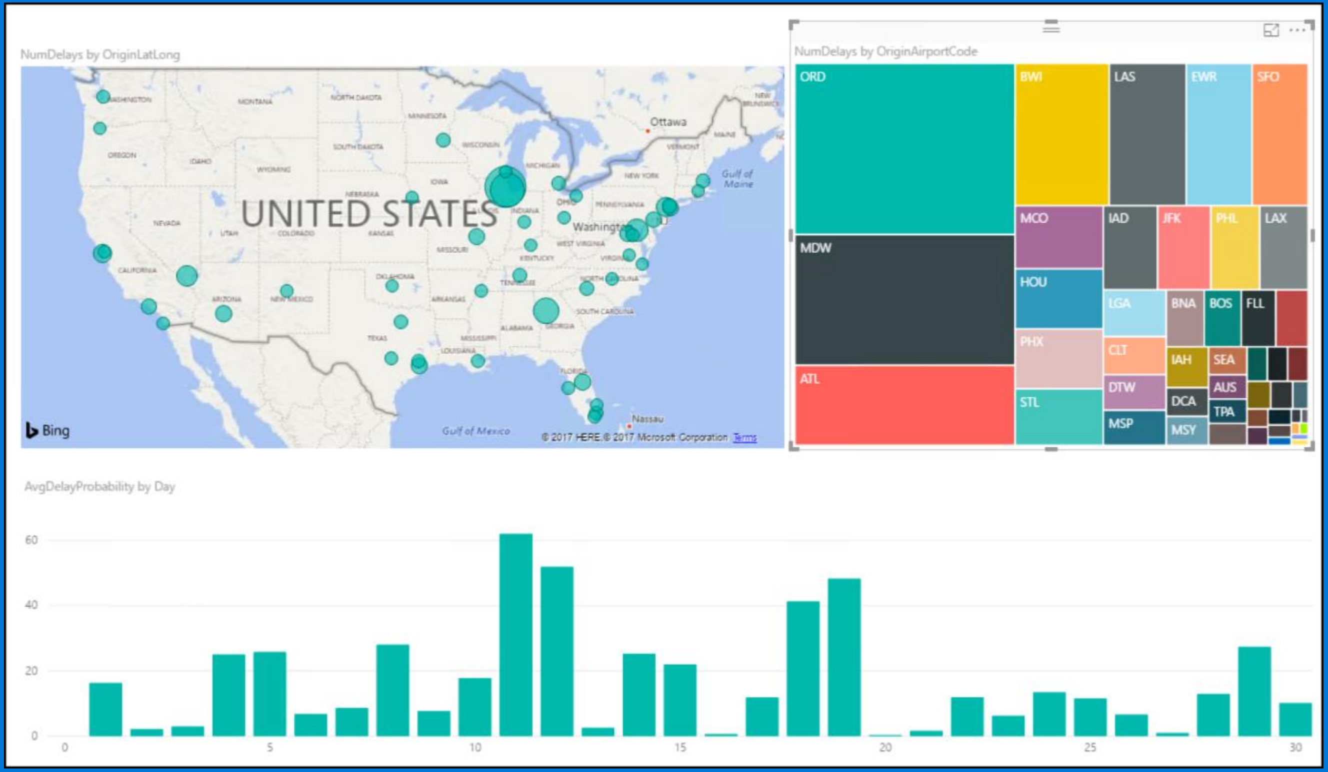
Operationalizing machine learning



Preferred solution

Visualization and reporting

- Power BI is a good option
- Direct Query against Spark Hive tables.
- Use map visualization



Preferred solution

Visualization and reporting

- Use Query Editor component of the Power BI Desktop, then upload to Power BI service.
- Create content pack with Power BI
- Restrict access in Azure AD

Customer objections

- Does Azure Machine Learning require a PhD in statistics?
- How long does it take to create and operationalize a machine learning model?
- Can operationalized ML models be flexible in the inputs they support?

Customer objections

- What are the options for running SQL on Hadoop solutions in Azure?
- Does Azure offer anything to speed up querying files in HDFS?
- How can we identify, monitor, and protect PII data?

Customer objections

- Is Azure Data Lake a good fit for our PoC?
- Can access to our SQL DW be limited using Azure Active Directory?
- What data visualization tools are available on Azure? Can access to these be managed with Active Directory?

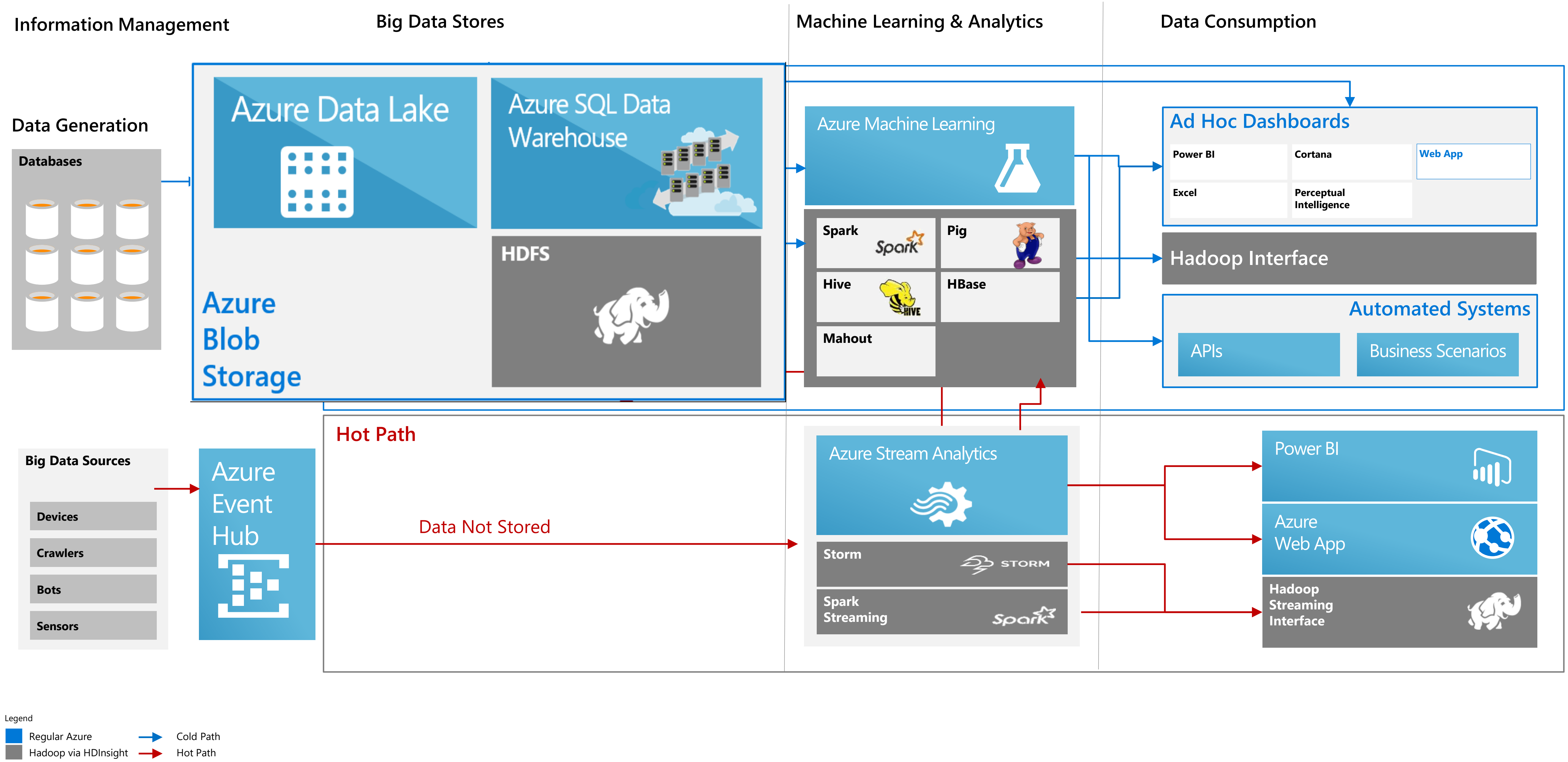
Customer quote

“We are flying into the future with Azure, helping our customers more aggressively schedule their travel, and optimize their non-travel time.”

- Jack Tradewinds, CIO of AdventureWorks Travel

AdventureWorks Travel

Azure Data Services



Information Management

Big Data Stores

Machine Learning & Analytics

Data Consumption

Data Generation

Databases

Azure Data Lake

Azure SQL Data Warehouse

HDFS

Azure Blob Storage

Azure Machine Learning

Spark

Pig

Hive

HBase

Mahout

Ad Hoc Dashboards

Power BI

Cortana

Web App

Excel

Perceptual Intelligence

Hadoop Interface

Automated Systems

APIs

Business Scenarios

Hot Path

Big Data Sources

Devices

Crawlers

Bots

Sensors

Azure Event Hub

Data Not Stored

Azure Stream Analytics

Storm

Spark Streaming

Power BI

Azure Web App

Hadoop Streaming Interface

Legend

🟦 Regular Azure

▨ Hadoop via HDInsight

→ Cold Path

→ Hot Path

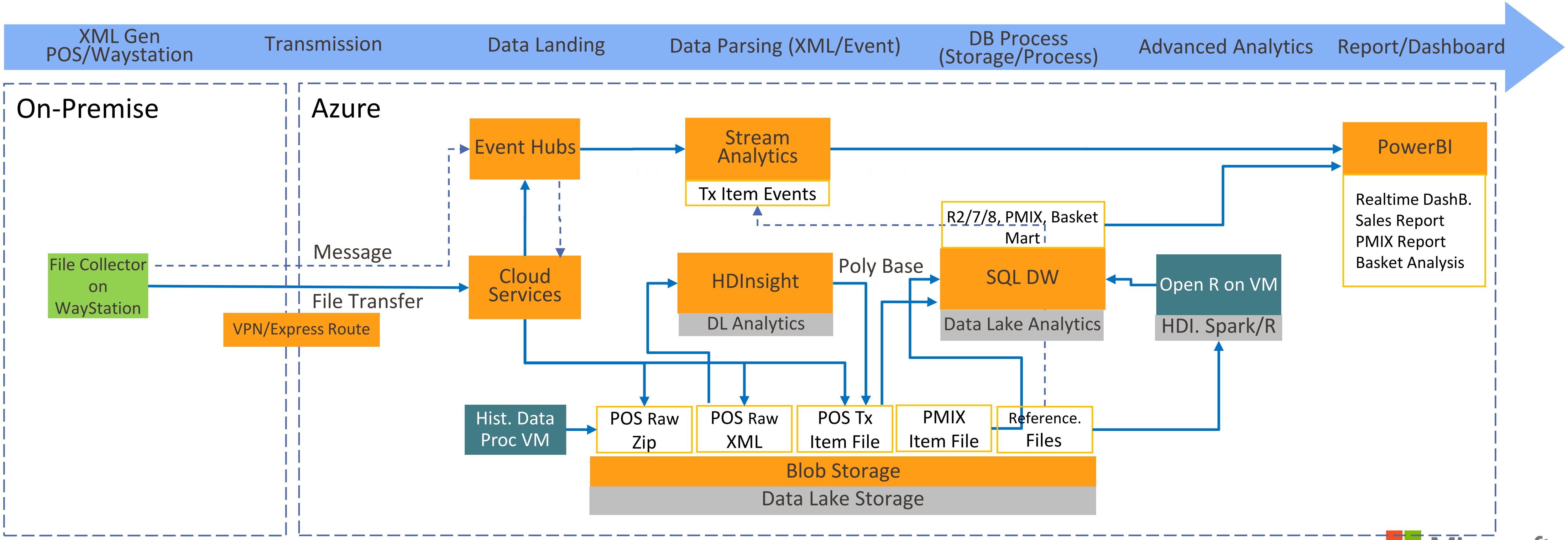
Solution Architecture

Cloud Product
IaaS
New Services
On-Premise

Data Onboarding
- Networking
- Protocol
- Update Approach

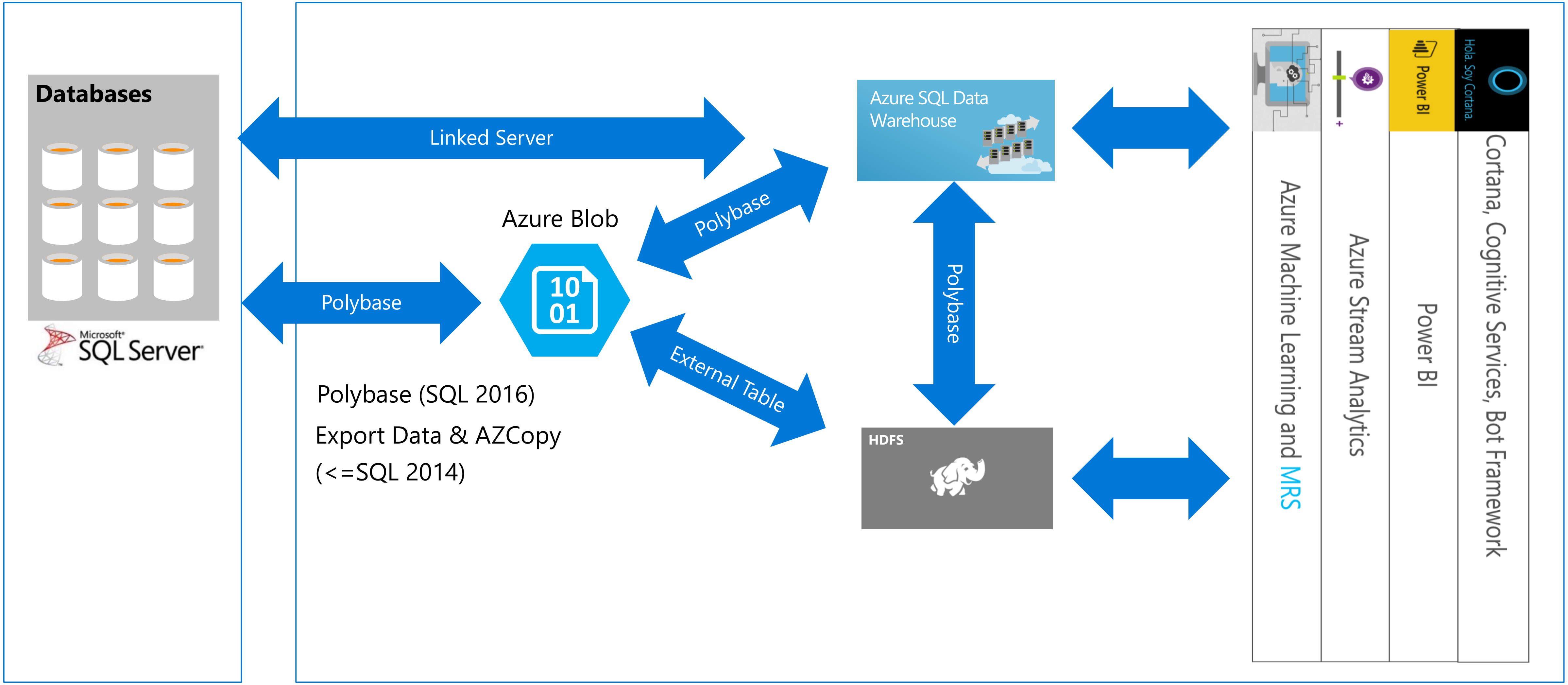
Data Storage
- On the fly
- Schema-less
- Schema-rich

Data Processing
- Transformation and Hybrid Data Warehousing
- Machine Learning
- Query Approach



Microsoft

Customer with SQL Server



Databases

Microsoft SQL Server

Linked Server

Polybase

Azure Blob

Polybase

External Table

Polybase (SQL 2016)

Export Data & AZCopy

(<=SQL 2014)

Azure SQL Data Warehouse

Polybase

HDFS

Azure Machine Learning and MRS

Azure Stream Analytics

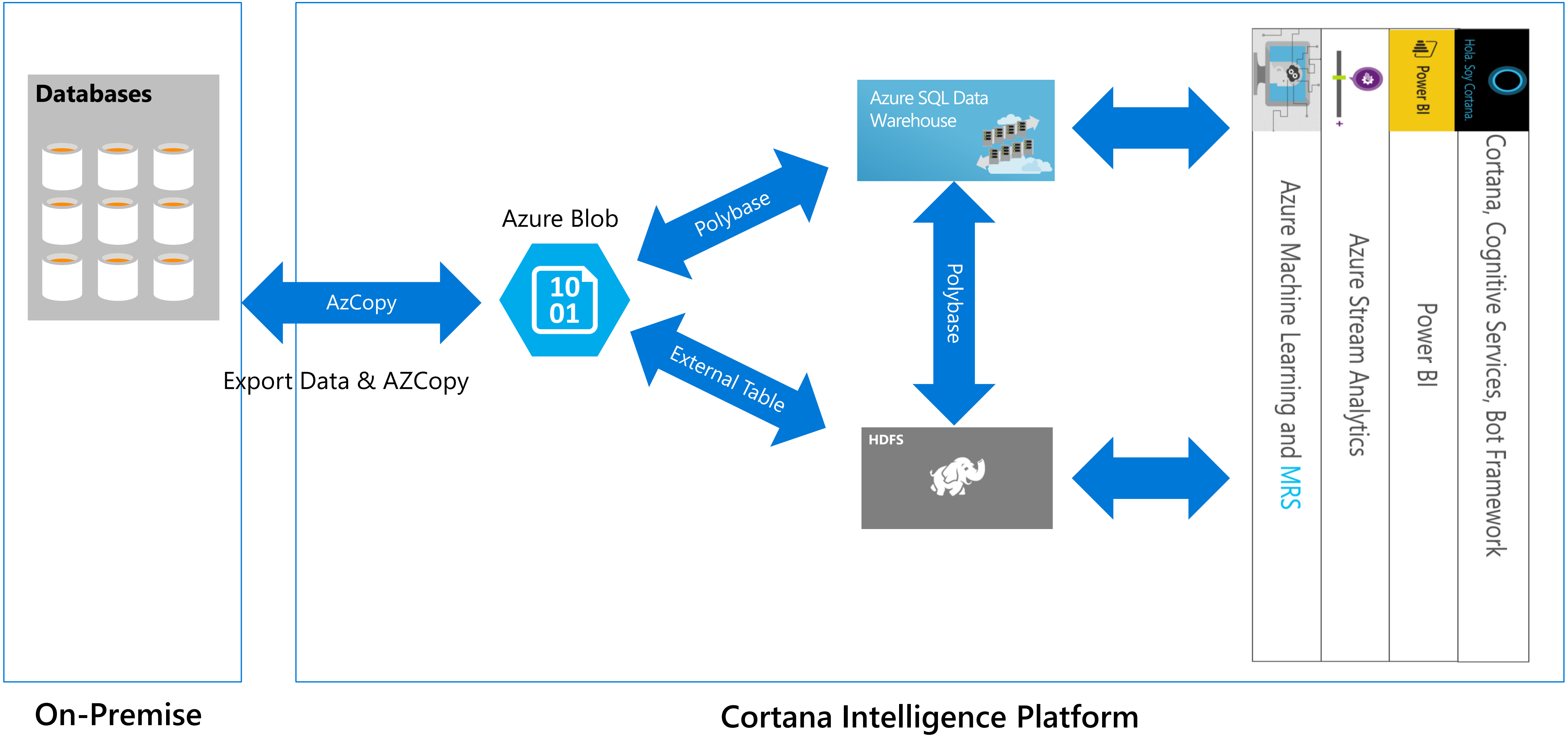
Power BI

Cortana, Cognitive Services, Bot Framework

On-Premise

Cortana Intelligence Platform

Customer with “Other” DB



Machine Learning in ML Studio

Anomaly Detection
- One-class Support Vector Machine
- Principal Component Analysis-based Anomaly Detection
- Time Series Anomaly Detection*

Classification
- Two-class Classification
  - Averaged Perceptron
  - Bayes Point Machine
  - Boosted Decision Tree
  - Decision Forest
  - Decision Jungle
  - Logistic Regression
  - Neural Network
  - Support Vector Machine
- Multi-class Classification
  - Decision Forest
  - Decision Jungle
  - Logistic Regression
  - Neural Network
  - One-vs-all

Clustering
- K-means Clustering

Recommendation
- Matchbox Recommender

Regression
- Bayesian Linear Regression
- Boosted Decision Tree
- Decision Forest
- Fast Forest Quantile Regression
- Linear Regression
- Neural Network Regression
- Ordinal Regression
- Poisson Regression

Statistical Functions
- Descriptive Statistics
- Hypothesis Testing T-Test
- Linear Correlation
- Probability Function Evaluation

Text Analytics
- Feature Hashing
- Named Entity Recognition
- Vowpal Wabbit

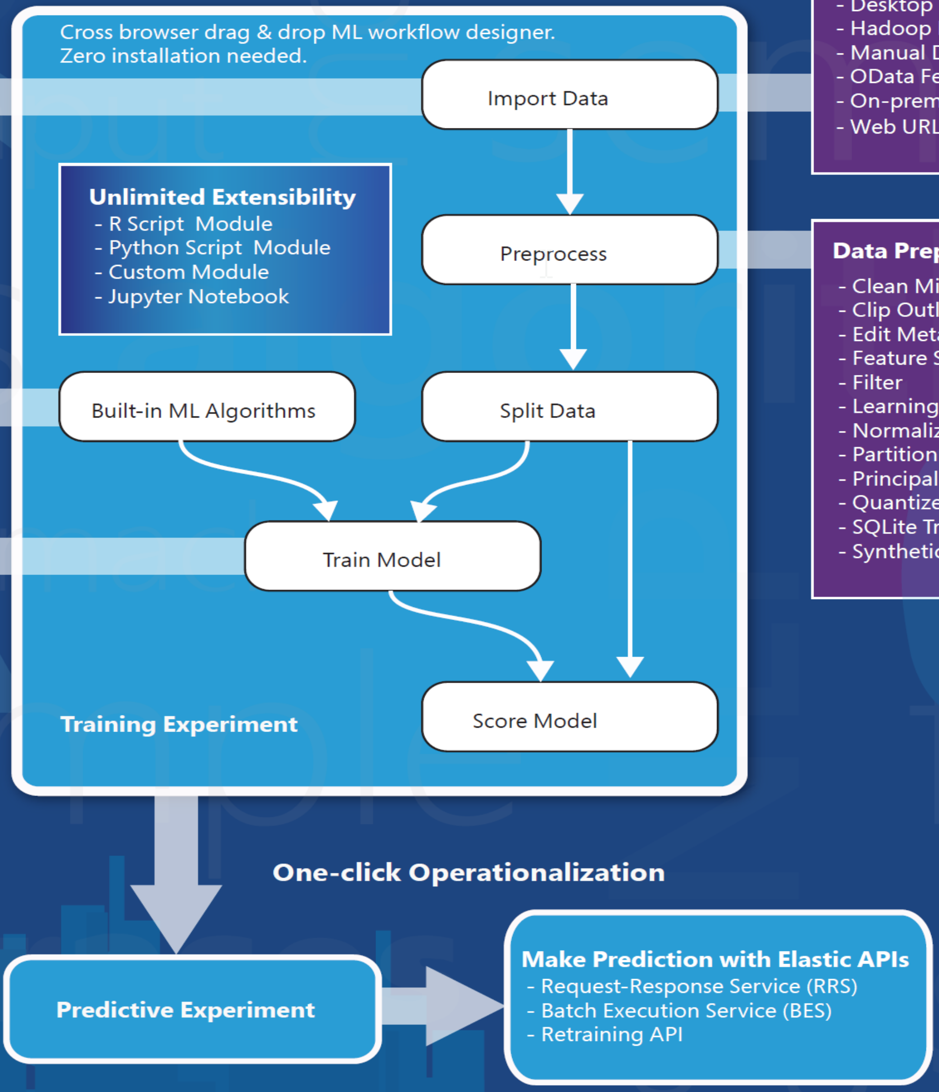
Computer Vision
- OpenCV Library

https://studio.azureml.net

Guest Access Workspace:	Free trial access without logging in.
Free Workspace:	Free persisted access, no Azure subscription needed.
Standard Workspace:	Full access with SLA under an Azure subscription.

Data/Model Visualization
- Scatterplots
- Bar Charts
- Box plots
- Histogram
- R and Python Plotting Libraries
- REPL with Jupyter Notebook
- ROC, Precision/Recall, Lift
- Confusion Matrix
- Decision Tree*

Training
- Cross Validation
- Retraining
- Parameter Sweep



Cross browser drag & drop ML workflow designer.
Zero installation needed.

Import Data

Preprocess

Split Data

Built-in ML Algorithms

Train Model

Score Model

Unlimited Extensibility
- R Script Module
- Python Script Module
- Custom Module
- Jupyter Notebook

Training Experiment

One-click Operationalization

Predictive Experiment

Make Prediction with Elastic APIs
- Request-Response Service (RRS)
- Batch Execution Service (BES)
- Retraining API

Data Source
- Azure Blob Storage
- Azure SQL DB
- Azure SQL DW*
- Azure Table
- Desktop Direct Upload
- Hadoop Hive Query
- Manual Data Entry
- OData Feed
- On-prem SQL Server*
- Web URL (HTTP)

Data Format
- ARFF
- CSV
- SVMLight
- TSV
- Excel
- ZIP

Data Preparation
- Clean Missing Data
- Clip Outliers
- Edit Metadata
- Feature Selection
- Filter
- Learning with Counts
- Normalize Data
- Partition and Sample
- Principal Component Analysis
- Quantize Data
- SQLite Transformation
- Synthetic Minority Oversampling Technique

Enterprise Grade Cloud Service
- SLA: 99.95% Guaranteed Up-time
- Azure AD Authentication
- Compute at Large Scale
- Multi-geo Availability
- Regulatory Compliance*

Community
- Gallery (http://gallery.azureml.net)
- Samples & Templates
- Workspace Sharing and Collaboration
- Live Chat & MSDN Forum Support

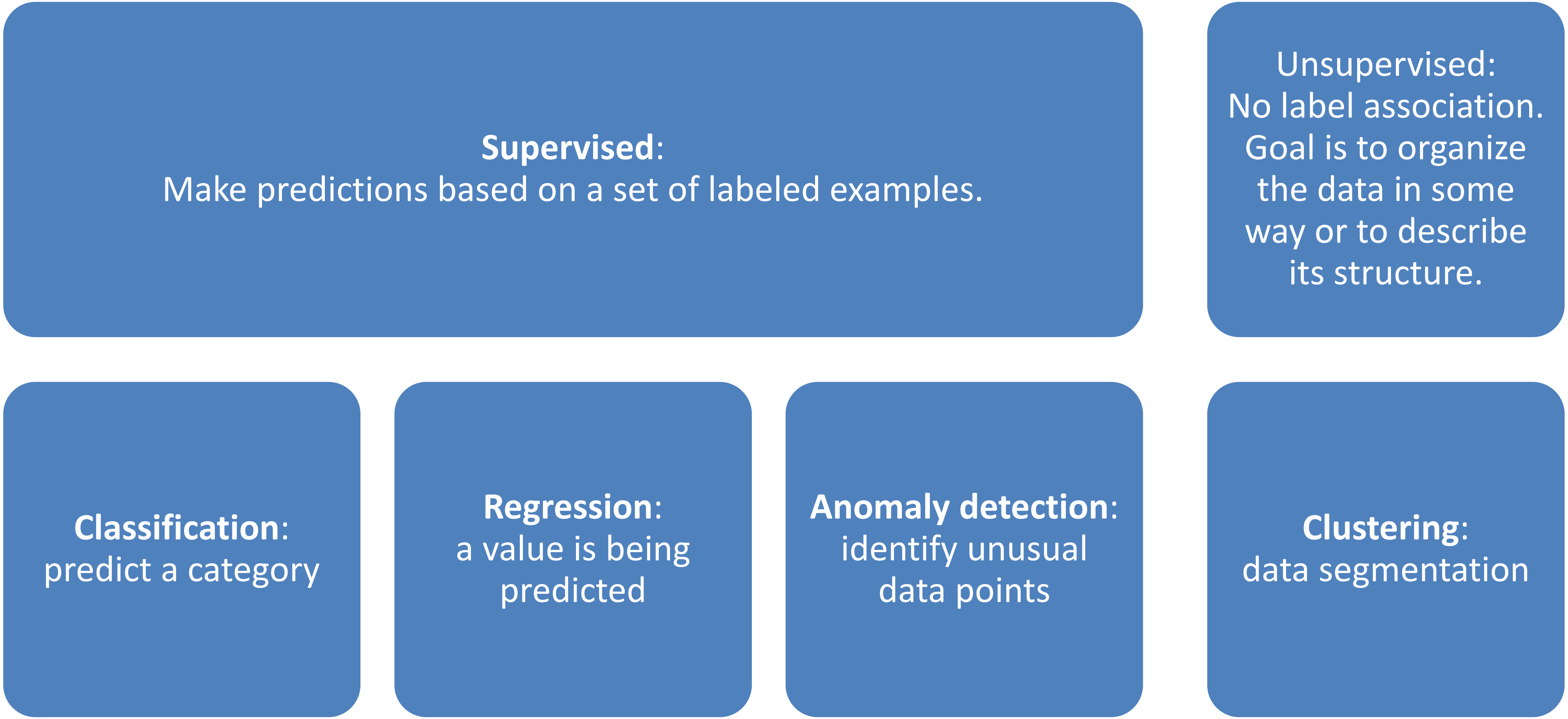
* Feature Coming Soon

Azure Machine Learning Studio Capabilities Overview

Microsoft

Created by the Azure Machine Learning Team   Email: AzurePoster@microsoft.com   Download this poster: http://aka.ms/MLStudioOverview

Select model type based on desired algorithm



Supervised:
Make predictions based on a set of labeled examples.

Unsupervised:
No label association.
Goal is to organize
the data in some
way or to describe
its structure.

Classification:
predict a category

Regression:
a value is being
predicted

Anomaly detection:
identify unusual
data points

Clustering:
data segmentation

Microsoft