Microsoft

//build/

Microsoft

L705

Data CodeLabs Module 2 – High-Scale Data Processing in Azure

Romit Girdhar
Software Engineer – Developer Experience Group

#Build2016

Key Takeaways

Understand how to process Big data in Azure

- Analyze terabytes and petabytes of data to identify patterns, understand your audience better or perform complex computations to improve your business.

Learn about the tools available to process your data

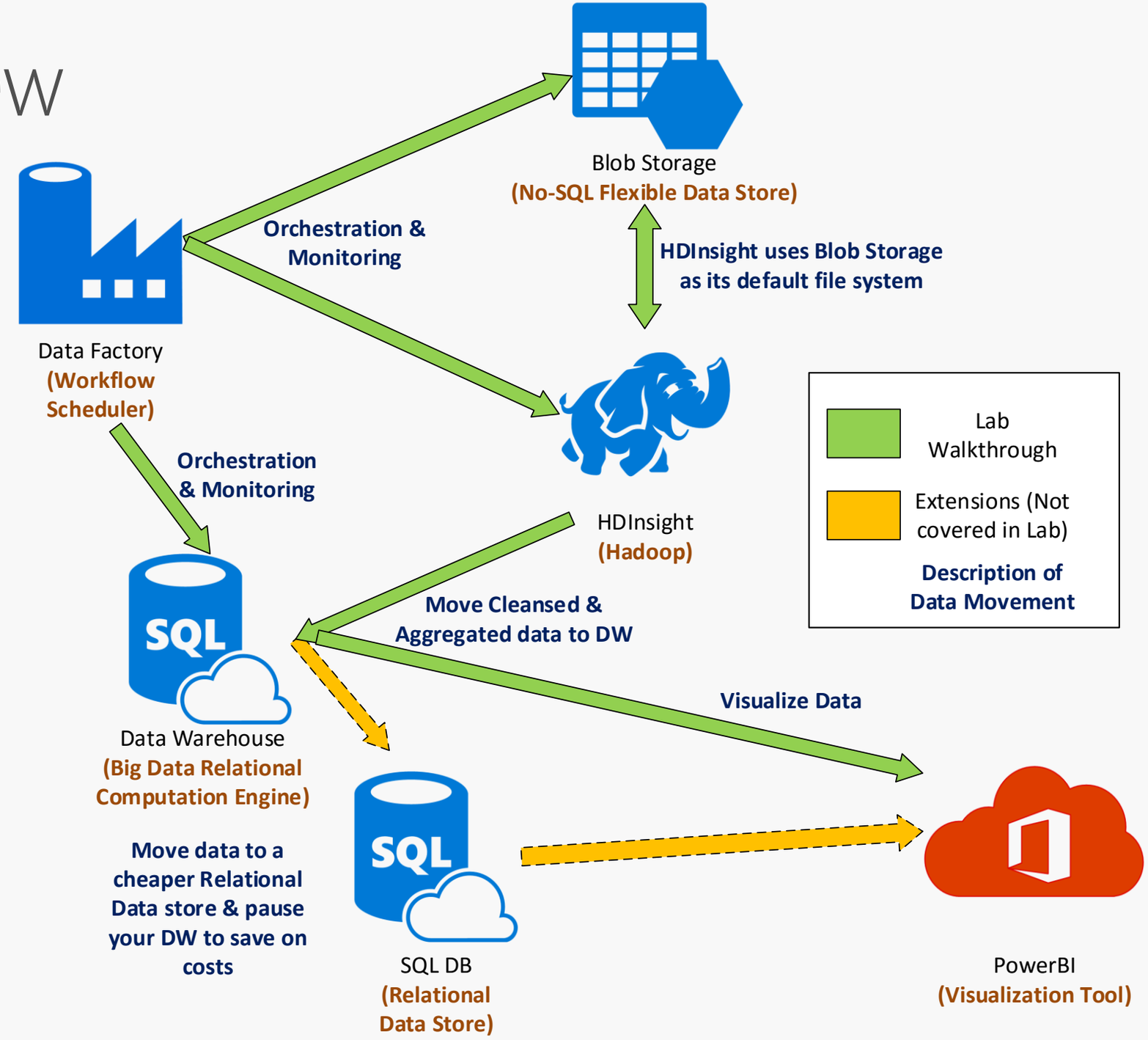
- From HDInsight (Apache Hadoop) to Azure SQL Data Warehouse, learn about the different tools available for you to process your data.

Orchestrate your Data Processing Workloads

- Using Azure Data Factory

Solution Overview

- Data stored in a cheap No-SQL storage gets picked up by your compute engine.
- The aggregated data gets moved to a "staging" data store
- PowerBI (or any other visualization tool) is used for visualization of data stored in the "staging" store.



Run the Setup.cmd

- Go through Steps 1, 2 & 3

- Open another Setup.cmd
- Go through Step #4

- 1st Three rows: East US
- 2nd Three rows: Central US
- 3rd Three rows: South Central US

(Alternatively) Create Azure HDInsight Cluster

- Go to the Azure Portal -> https://portal.azure.com
- Click on New -> Data & Analytics -> HDInsight
  - Cluster Name: **hdicluster-{SuffixNumber}**
  - Select Cluster Type -> OS: **Windows/Linux** & Version: **2.7.0**
  - Credentials -> Username: **admin** ; Password: **P@ssword123** (or something else…)
  - Data Source -> Create a New Storage Account
    - Location: **West US**
  - (Optional) Configure your Node VM sizes to optimize cost

(Alternatively) Create SQL DW Cluster

- From the Azure Portal
- Click New -> Data & Storage -> Data Warehouse
  - SQL DW Server Name: **buildlab-{SuffixNumber}**
  - SQL DW Name: **buildlab-dw**
  - Username: **dwadmin**
  - Password: **P@ssword123** (or anything else you'd like)
  - Location: **West US**

Azure Data Factory

- Create an Azure Data Factory
- Go to New -> Data & Analytics -> Data Factory
  - Name: **datalab-factory-{Suffix Number}**
  - Location: **West US**

Azure Data Factory

- Create Linked Services
  - Azure Storage Linked Service
  - HDInsight Linked Service

Azure Data Factory

- Create Datasets
  - Azure Storage Raw Dataset
  - Azure Storage Dummy Dataset

Azure Data Factory

- Create Pipeline
  - Adding Partitions for Hive

HDInsight

- Enable SSH on your Cluster through the Azure Portal
- Ssh into your cluster
- Putty is located under C:\CodeLabs-Data\putty.exe
- Enter the following Link in Putty:

  {Cluster Name}-ssh.azurehdinsight.net
- Open the script '**createtables.hql**' located in the '**Scripts**' folder
- Replace the Hive variables with your storage account name and execute!

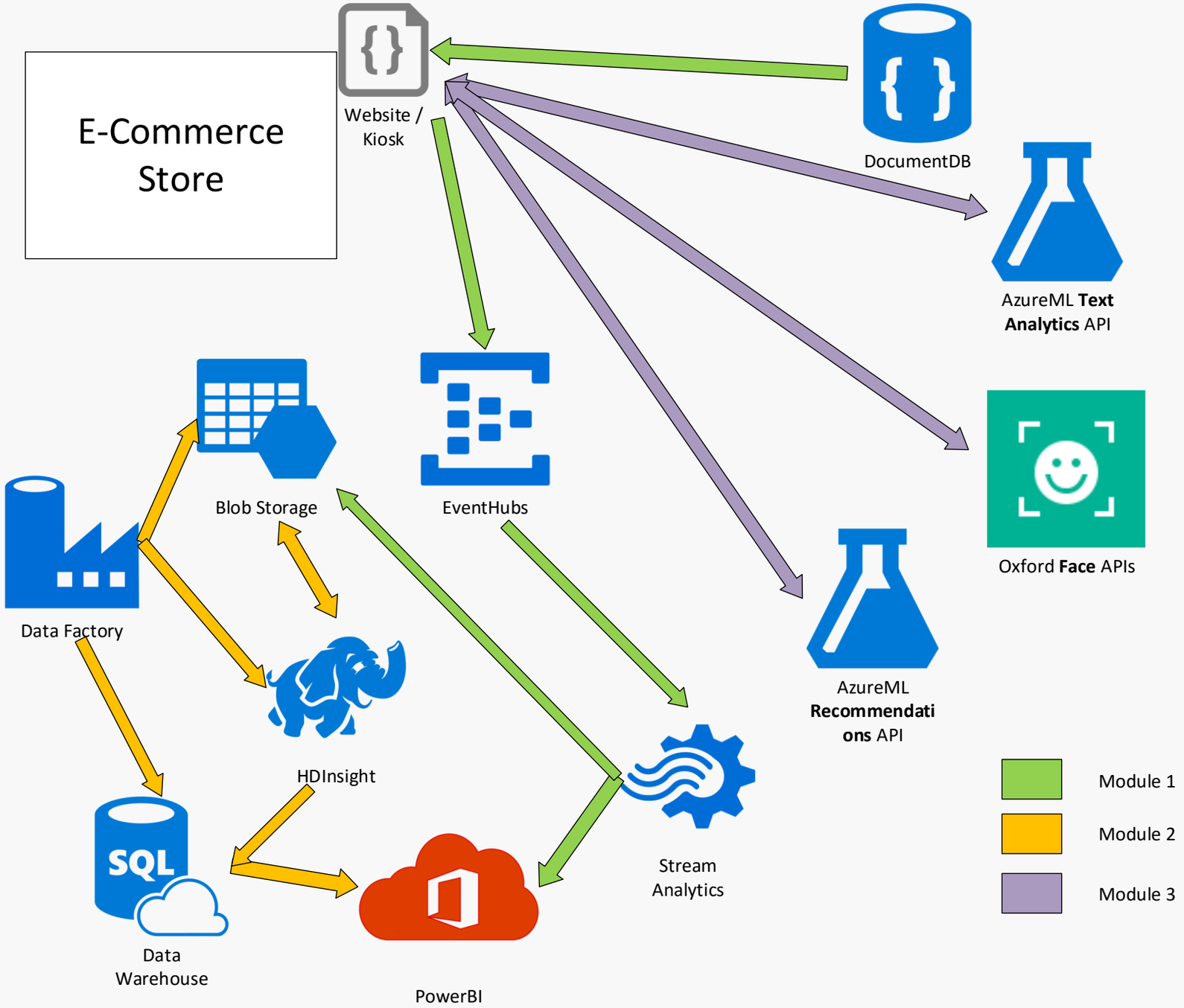
Data Factory Automation

- Open Command Prompt (Start -> Type “cmd”)
- Go to: C:\CodeLabs-Data\Tools\ADFSetup
- Type the following:

ADFSetup.exe <SubscriptionID> <Resource Group Name> <DataFactoryName> <part1/part2/all>

- Follow the Prompts…

Architecture Across The Data Modules



Please Complete An Evaluation Form
Your input is important!

Required Slide
*delete this box when your slide is finalized

or