데이타솔루션 Microsoft

Azure Databricks

A Technical Overview

Scale Analytics Hands on Lab Day 2

01 | What is Azure Databricks?

Azure Databricks

A fast, easy and collaborative Apache® Spark™ based analytics platform optimized for Azure

- Apache Spark 창립자와 공동으로 설계
- 원 클릭 설정, 간소화 된 워크 플로우, 단일 청구서
- 데이터 과학자, 데이터 엔지니어 및 비즈니스 분석가 간의 협업을 가능하게하는 대화식 작업 공간.
- Azure 서비스 (Power BI, SQL DW, Cosmos DB, Blob Storage)와의 기본 통합
- 엔터프라이즈 급 Azure 보안 (Active Directory 통합, Compliance, 엔터프라이즈 레벨의 SLA)

02 | Azure Databricks

- **Azure Databricks는 Azure의 자사 서비스입니다.**
  - 다른 클라우드와 달리 Azure Marketplace 또는 타사 호스팅 서비스가 아닙니다.
- **Azure Databricks는 Azure 서비스와 완벽하게 통합됩니다.**
  - 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
- Databricks를 사용하여 별도의 계정을 만들 필요가 없습니다.

03 | Data Scientists & Data Engineers

데이터 과학자

패턴분석 및 향후예측을 위해
데이터 분석

PAIN POINTS/CONCERNS

- Often spends too much time on accessing/ ingesting data. Exploration at scale is difficult

Azure Databricks Opportunity

- Get to tool in their hands ASAP, it increases their productivity
- Azure + Spark + Databricks = great resume builder
- Can be your best champion
- Be careful of devs & data engineers “rebranding” as data scientists
- Trouble accessing budget, focus on finding value

데이터 엔지니어

ETL / Cleansing을 통해
비기술적인 최종 사용자에게 원시 데이터를
사용가능한 형태로 전환

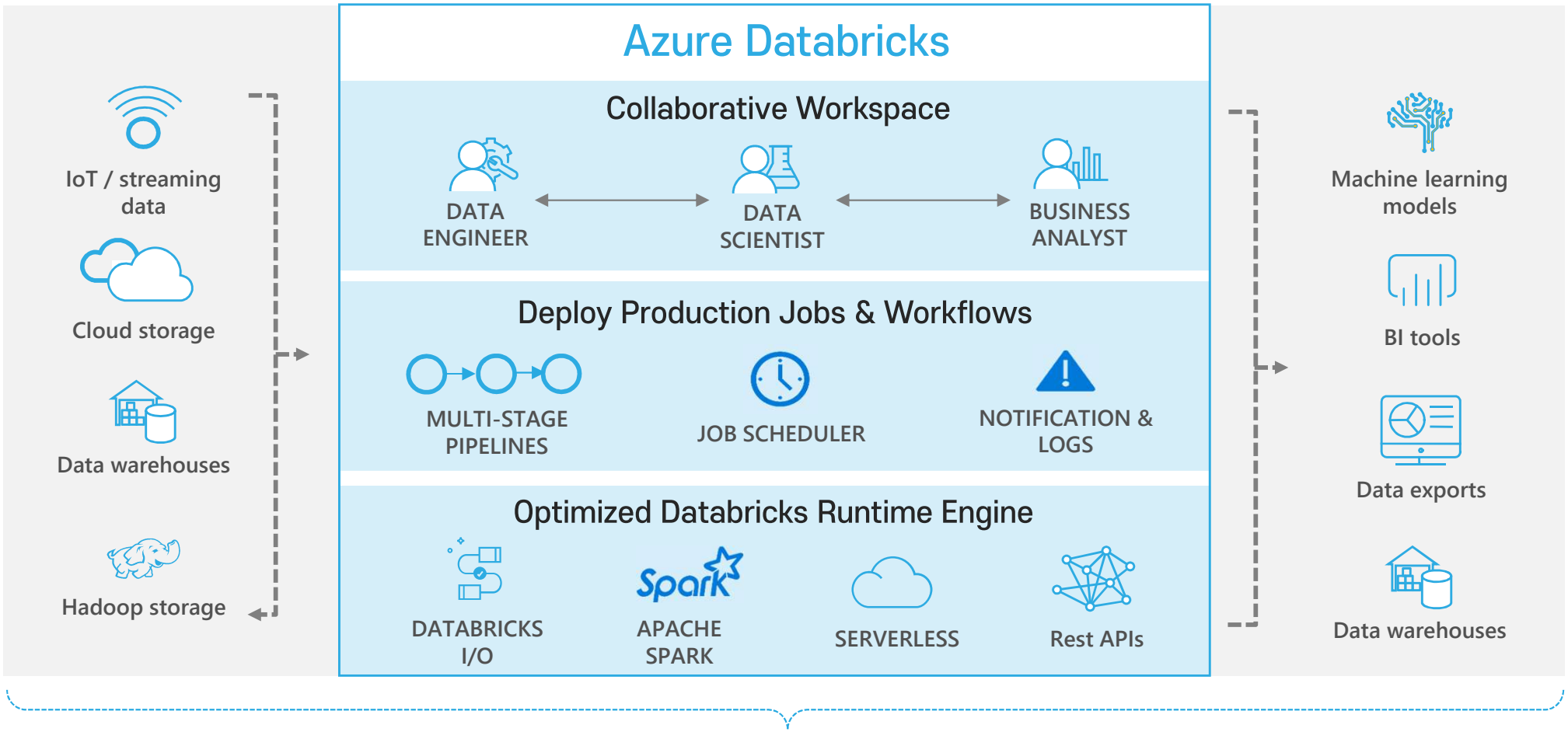
PAIN POINTS/CONCERNS

- Difficult to do fast and reliably enough to support the business when dealing with scale, and variety of data sources and types. Painful to access and ETL data.

Azure Databricks Opportunity

- Easier and faster data access and ETL, cost effective and zero-maintenance infrastructure
- Very careful about production grade deployments
- They want programmable control of the platform
- Focus on APIs, performance and reliability
- Can be very cheap, focus on finding value.

04 | Azure Databricks



Enhance Productivity | Build on secure & trusted cloud | Scale without limits

1 Chapter

Azure Databricks Core Concepts

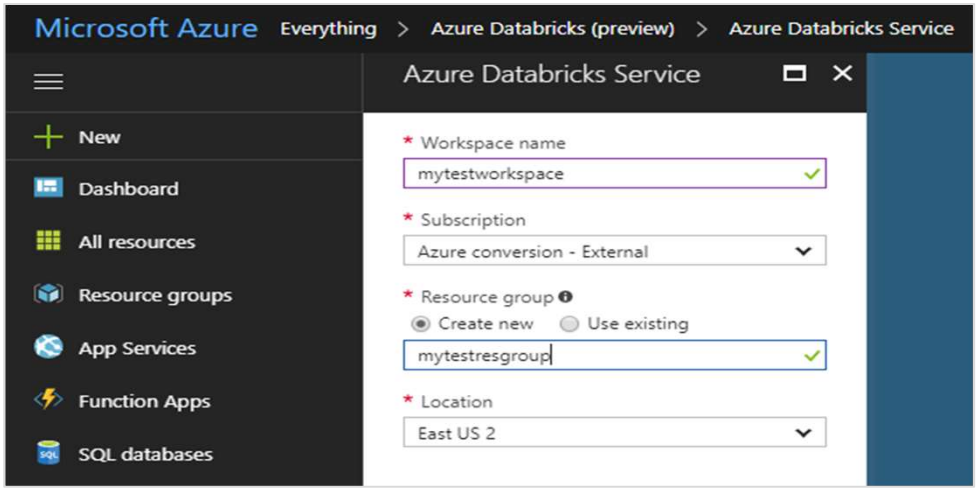
데이타솔루션 Microsoft

01 | Provisioning Azure Databricks WORKSPACE

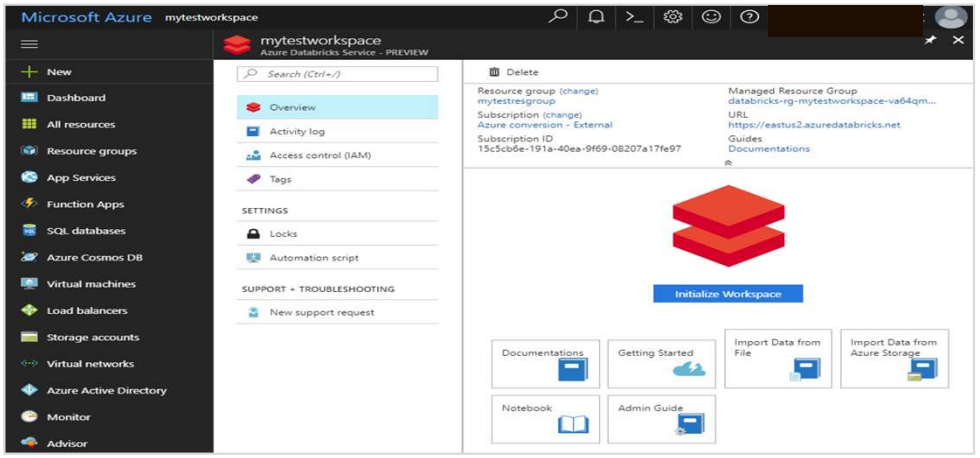
- Azure Databricks is provisioned directly from the Azure Portal like any other Azure service
  - In contrast, with other clouds, it has to be provisioned through the Databricks portal.
  - With Azure Databricks, the Azure Portal offers a unified portal to provision and administer Azure Databricks as well as other Azure services.

- Any Azure user with the appropriate subscription and authorization can provision Azure Databricks service*.
  - There is no need for a separate Databricks account

** During the current preview phase, the subscription has to be whitelisted.*

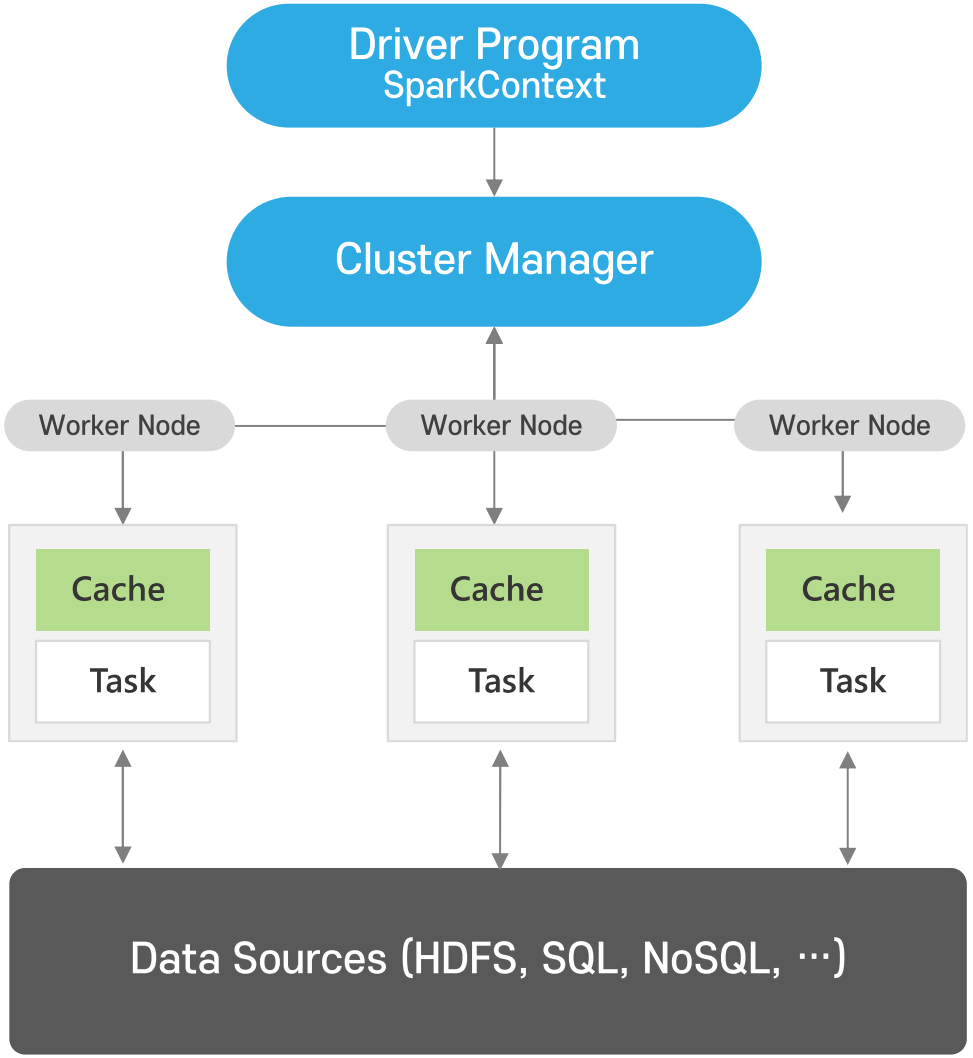


▲ *Provisioning the Azure Databricks Service*



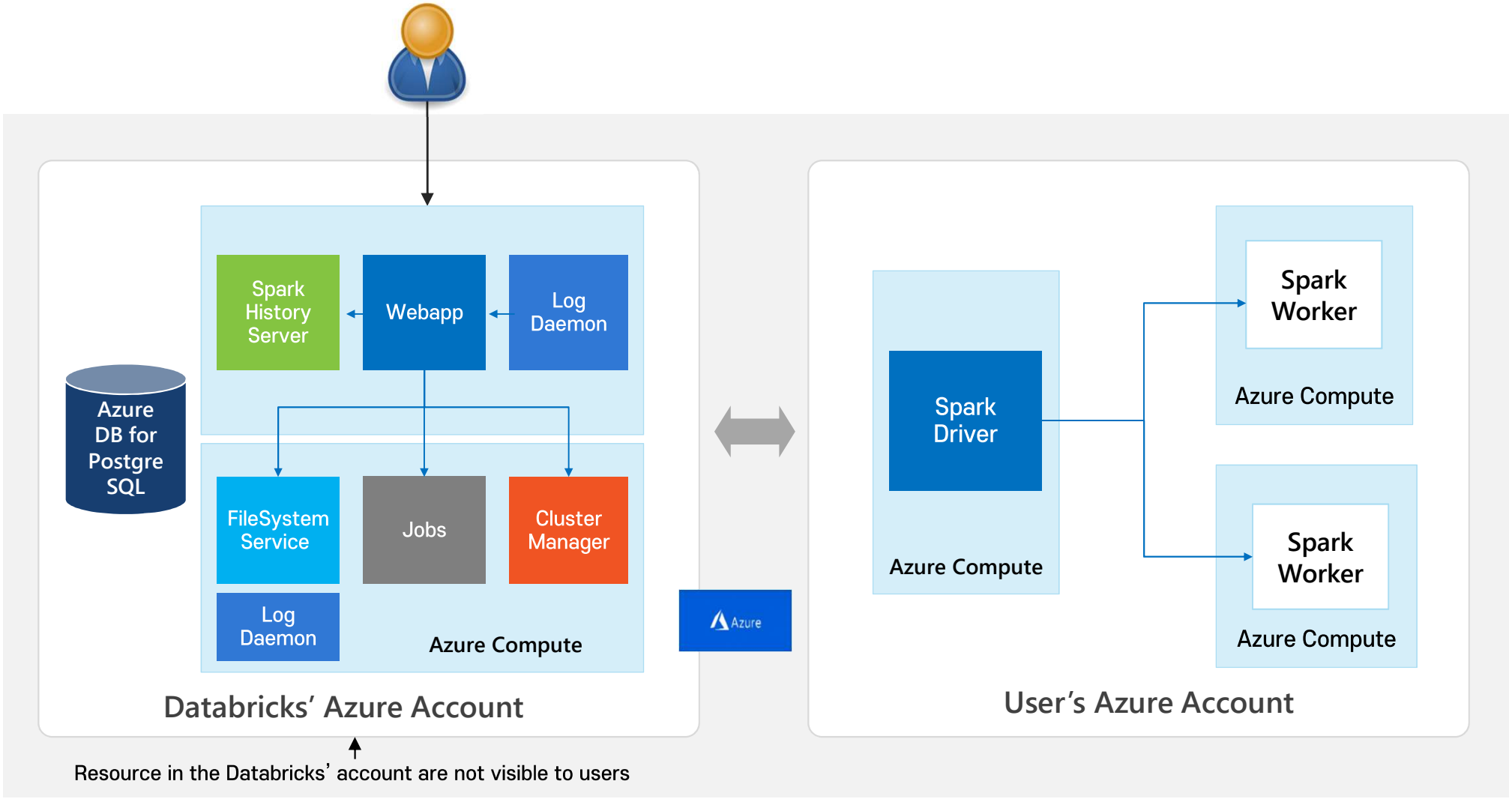
▲ *After provisioning the is complete*

02 | General Spark Cluster Architecture

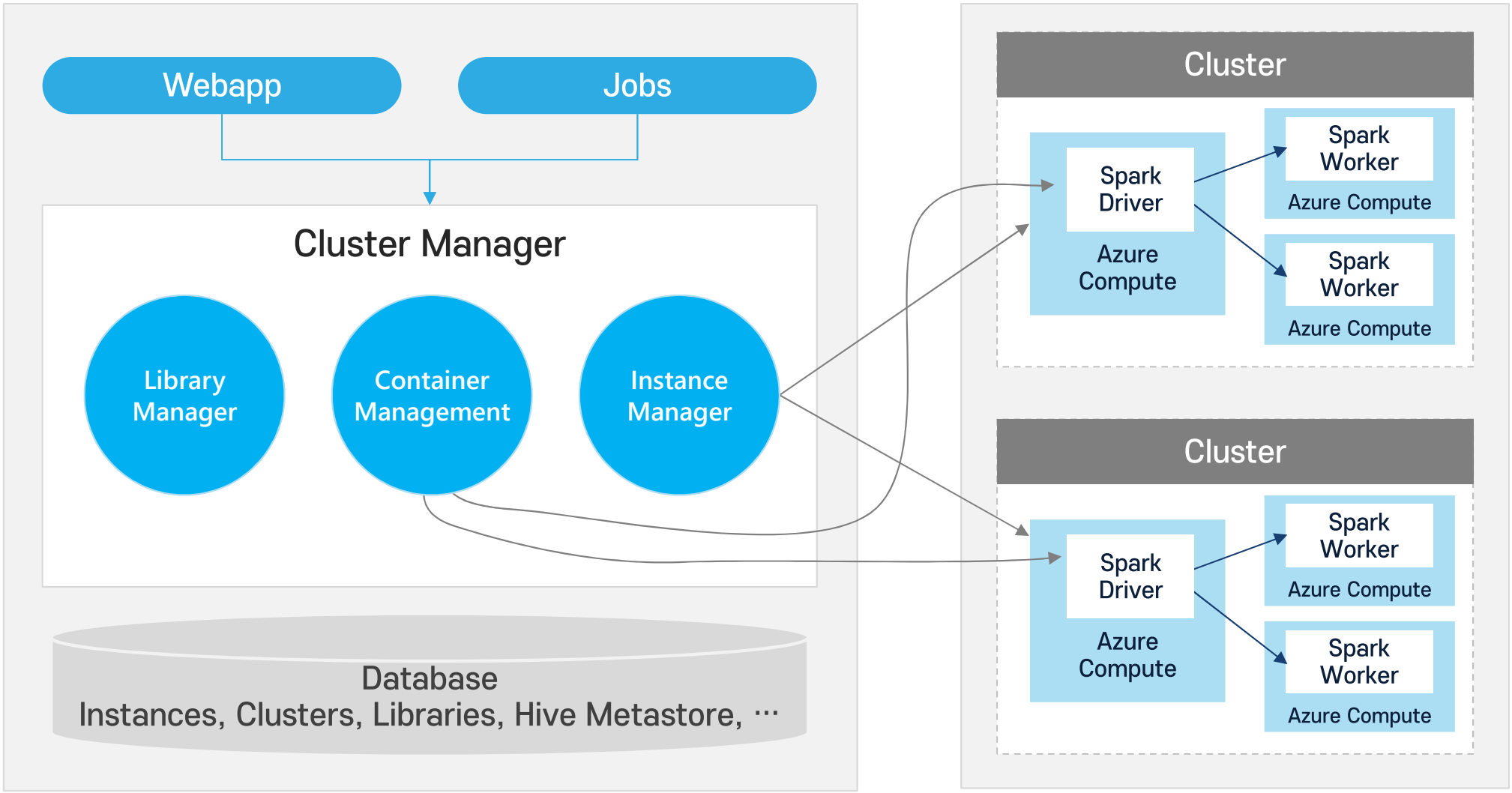


- 'Driver' runs the user's 'main' function and executes the various parallel operations on the worker nodes.
- The results of the operations are collected by the driver
- The worker nodes read and write data from/ to Data Sources including HDFS.
- Worker node also cache transformed data in memory as RDDs (Resilient Data Sets).
- Worker nodes and the Driver Node execute as VMs in public clouds (AWS, Google and Azure).

03 | Azure Databricks Cluster architecture

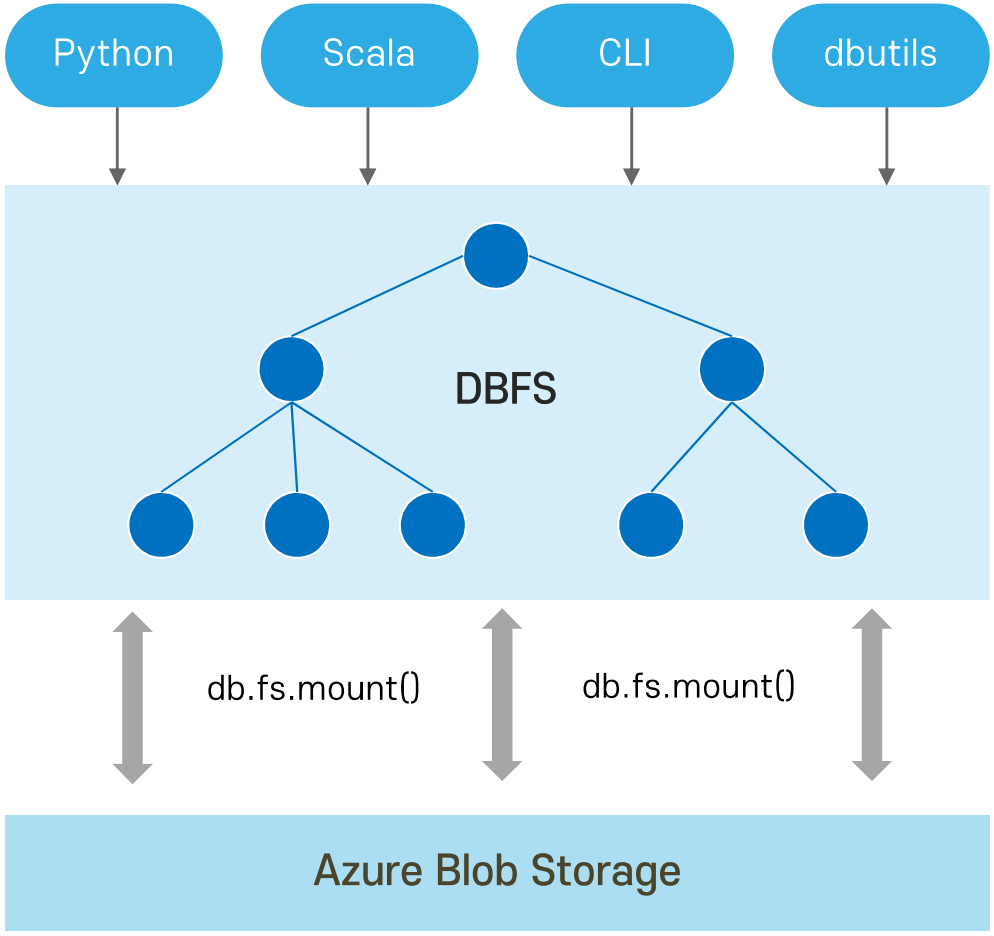


04 | Cluster Manager Architecture



05 | Databricks File System (DBFS)

Is a distributed File System (DBFS) that is a layer over Azure Blob Storage



- Azure Storage buckets can be mounted in DBFS so that users can directly access them without specifying the storage keys
- DBFS mounts are created using dbutils.fs.mount()
- Azure Storage data can be cached locally on the SSD of the worker nodes
- Available in both Python and Scala and accessible via a DBFS CLI
- Data persist in Azure Blob Storage – is not lost even after cluster termination
- Comes pre-installed on Spark clusters in Databricks

데이타솔루션 Microsoft

Azure Machine Learning service:

A Technical Overview

Scale Analytics Hands on Lab Day 2

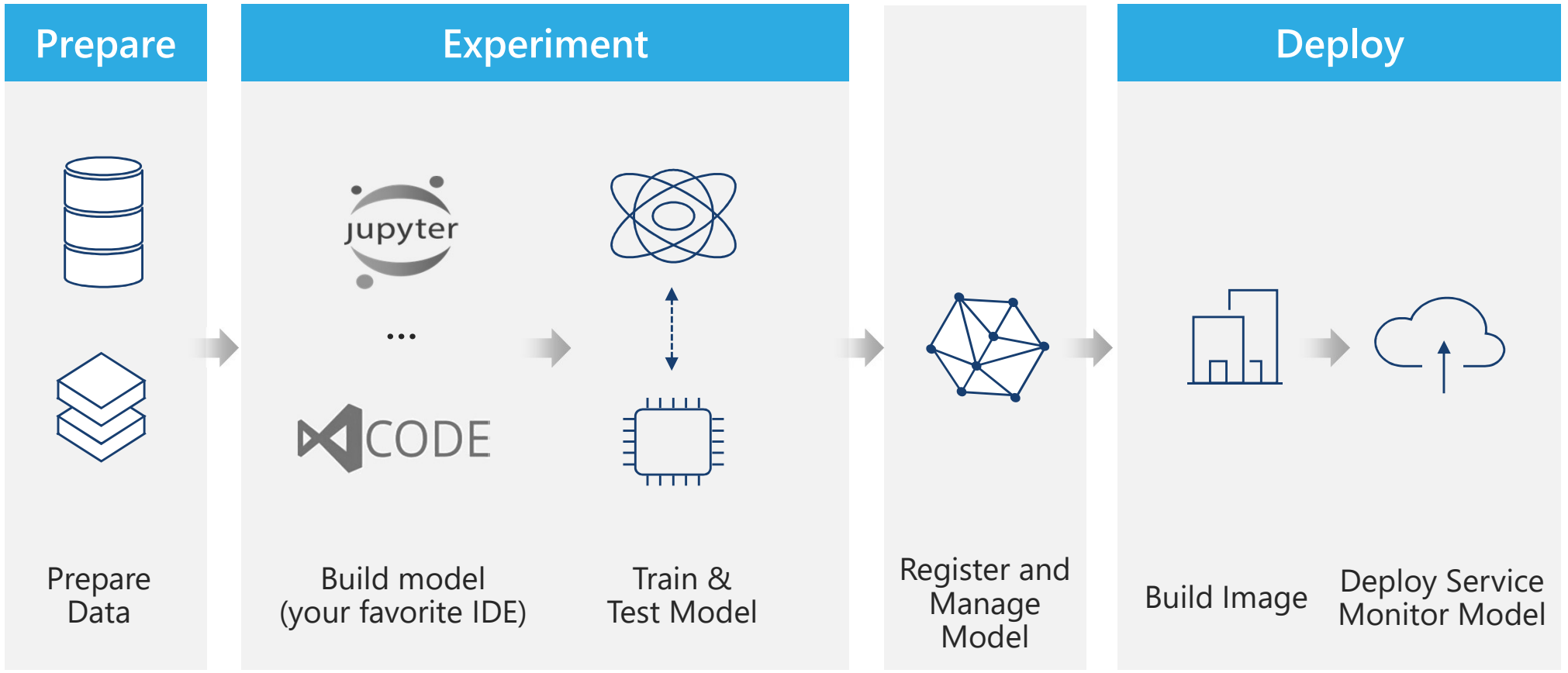
1 Chapter

Requirements of an advanced ML Platform

데이타솔루션 Microsoft

01 | Machine Learning

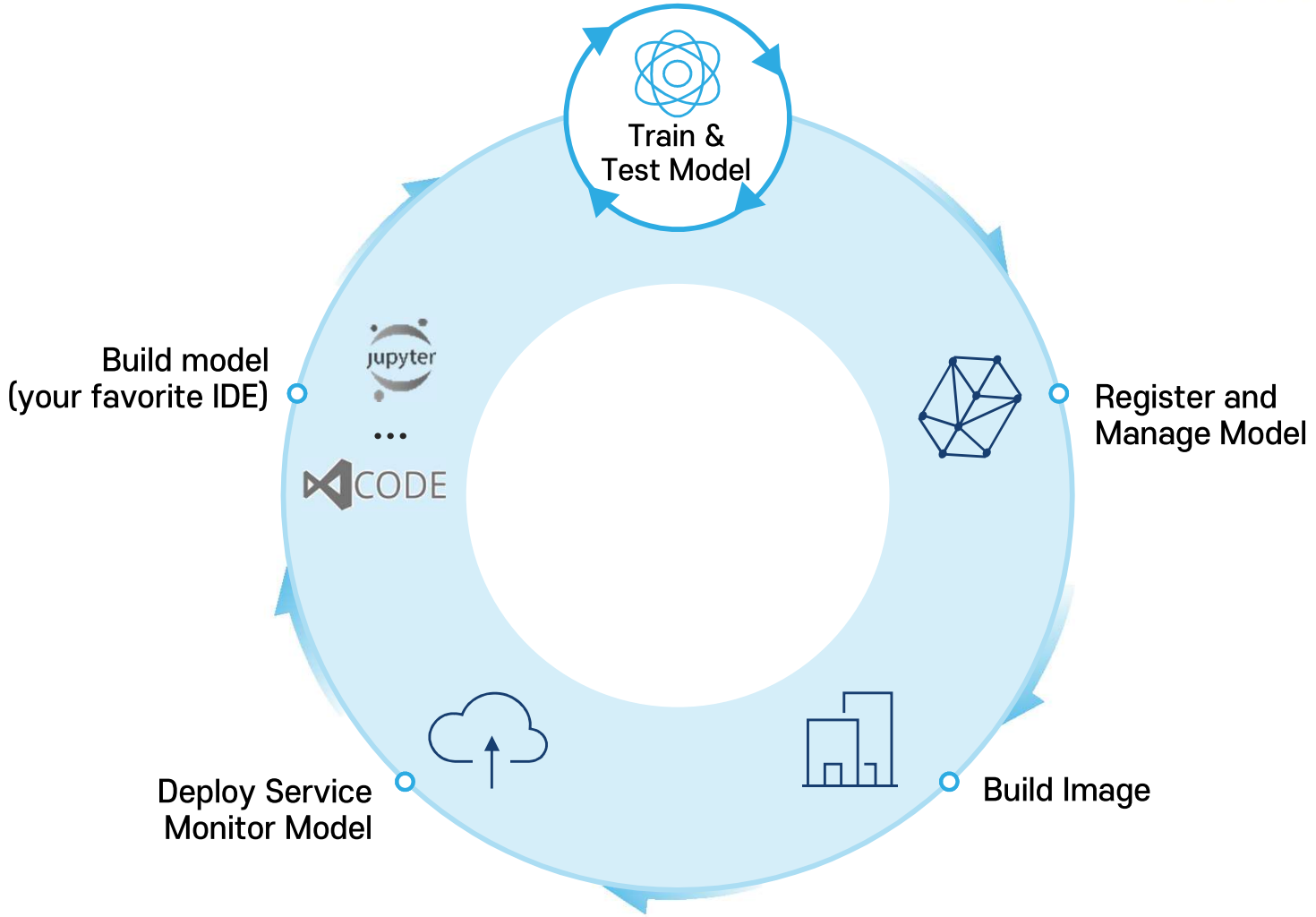
Typical E2E Process



02 | DevOps loop for data science

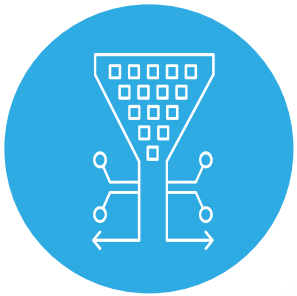
Prepare

Prepare Data



03 | Data Preparation

Requirements



1 Multiple Data Sources

- SQL 및 NoSQL 데이터베이스, 파일 시스템, 네트워크 연결 저장소 및 클라우드 저장소 (예 : Azure Blob Storage) 및 HDFS.

2 Multiple Formats

- Binary, text, CSV, TS, ARFF, etc.

3 Cleansing

- NULL values, outliers, out-of-range values, duplicate rows를 감지하고 수정합니다.

4 Transformation

- 일반 데이터 변환 (변환 유형) 및 ML 특정 변환 (인덱싱, 인코딩, 벡터 어셈블링, 벡터 정규화, Binning, 정규화 및 분류).

04 | Model Building

Requirements

1 Choice of algorithms

2 Choice of language

- Python

3 Choice of development tools

- Jupyter, PyCharm 및 Spark Notebook과 같은 브라우저 기반 REPL 지향 노트북.
- Visual Studio 및 R-Studio for R 개발과 같은 데스크탑 IDE.

4 Local Testing

- To verify correctness before submitting to a more powerful (and expensive) training infrastructure.

05 | Model Training

Requirements

1 Powerful Compute Environment

- 스케일 업 VM, 자동 스케일링 스케일 아웃 클러스터를 선택 할수 있어야 합니다.

2 Preconfigured

- 컴퓨팅 환경은 모든 정확한 버전의 ML 프레임 워크, 라이브러리, 실행 파일 및 컨테이너 이미지로 사전에 설정되어야 합니다.

3 Job Management

- 데이터 과학자는 작업을 쉽게 시작, 중지, 모니터링 및 관리 할 수 있어야합니다

4 Automated Model and Parameter Selection

- 솔루션은 원하는 결과를 위해 자동으로 최상의 알고리즘과 해당하는 최고의 하이퍼 파라미터를 선택해야 합니다.

07 | Model Registration and Management

Requirements

1 Containerization

- 실행 환경에 배치 할 수 있도록 모델을 Docker 컨테이너로 자동 변환되어야 합니다.

2 Versioning

- A / B 테스트, 변경 롤백 등을 위해 모델에 버전 번호를 지정하고, 시간이 지남에 따라 변경 사항을 추적하고, 배포 할 특정 버전을 식별하고 검색합니다.

3 Model Repository

- 모델 저장 및 공유를 위해 CI / CD 파이프 라인에 통합 가능해야 합니다.

4 Track Experiments

- 감사를 위해 시간이 지남에 따른 변경 사항을 확인하고 팀 구성원 간 협업이 가능해야 합니다.

08 | Model Deployment

Requirements

1 Choice of Deployment Environments

- Single VM, Cluster of VMs, Spark Clusters, Hadoop Clusters, In the cloud, On-premises

2 Edge Deployment

- 이벤트 소스에 근접한 예측을 가능하게 하고 불필요한 데이터 전송을 피하며 더 빠른 응답을 가능하게 합니다.

3 Security

- 에지에 배포 된 경우에도 E2E 보안을 유지해야합니다. 안전한 인증 된 장치에만 모델을 배포하고 데이터를 전송해야합니다.

4 Monitoring

- 상태, 성능 및 보안 모니터링이 가능해야 합니다.

Azure Databricks | A Technical Overview

Thank you

데이타솔루션 Microsoft