Architecting a Data Analytics Service Both in the Public Cloud and in the On-Premise Private Cloud: ETL, BI, and Machine Learning

2019. 9. 25
Agenda

I. Who are we?

II. Trend of Data Analytics Platform

III. Suggesting an Architecture

IV. Offering for Data Analytics Service
   • Public Cloud
   • On Premise Cloud
   • Hybrid Cloud

V. Roadmap

VI. Demo
Who are we?

• SK Group is the second largest business group in South Korea.
• Composed of **95 subsidiary companies**
• Share the **SK brand name** and **Group's management culture**
• More than 100,000 employees across worldwide
• **SK holdings owns SK Companies**
• One of top three IT service providers in South Korea
• **Deal with all kind of data from all kind of industries**
  - In the SK Group: SK Innovation for petroleum, SK Telecom (26M subscribers), SK Broadband for IPTV, SK Hynix - the world's second largest semiconductor manufacturer, etc.
  - Outside of SK Group: Many other customers in Banks, Retails, Manufacturing, Insurance and so on
• **Developed our own data analytics platform**
Trend of Data Analytics Platform

Data Analytics Platforms require a number of features

- **Data Lake** to save all kind of Data
- **Data Catalog** can be used for finding the data from Data Lake
- Independent Analytics Environment with the *Sand Box*
- **Data Pipeline** for ETL and Operations
- Various **ML / DL Analytics Tools** should be provided
- **AutoML** for finding the best algorithms with optimized hyper parameters
- **Analytics Model Life Cycle Management**
- Flexible **Scalability** should be guaranteed for many analysis works
- **Collaboration** tool is required for Data Analysis
Trend of Data Analytics Platform

Conceptual Diagram for Data Analytics platform based on Cloud

Various Data Types
Structured Data
Unstructured Data

- Storing all kind of Data (Law, RDBMS, XML,
- Easy to Search
- Extracting Data

Data Lake

ETL with Data Catalog

Data Catalog

Scalability based on cloud

Real-Time Process
Batch Process

Inference Area
R
Python

Auto Scaling
Model Performance Monitoring

Sandbox for Data Analysis
ML Model
DL Model
AutoML

Sandbox for Computing
Suggesting Architecture - Overall

**PaaS Layer with Data APIs**

- Collection
- Machine Learning
- Real-Time Streaming
- Hadoop Batch
- Query
- Visualization
- Deep Learning
- Hadoop Provisioning

Docker Container orchestration, Multi-Tenancy, MSA, Authentication/Authorization, Job & Log Management(Spark/MapReduce/Tensorflow), Data & Model Life Cycle Managements

**IaaS Layer**

**Data Lake**

- Object Storage
- Hadoop Cluster

**Modeling & Development**

- AccuInsight+
- R, Python
- DL Modeler
- ML Modeler
- BigQL
- DHP
- Batch Pipeline
- Realtime Pipeline
- Cloud Search
- Data Insight
- R, Python Modeler
- Data Catalog

**Hadoop Cluster**

- R
- Apache Spark
- Docker

**GPU**

- PyTorch
- Keras
- TensorFlow

**Kubernetes**

**RANCHER**
Suggesting Architecture - Details

PaaS Layer with Data APIs

- Host A: Hadoop Provisioning
- Host B: Batch Pipeline, Real-Time Pipeline
- Host C: Real-Time Pipeline, Batch Pipeline
- Host D: Machine Learning

Hazelcast Cluster Manager

Workers

- BP, ML, RP
- HP, ML, RT
- HP, BP, ML

Shared Cache for End points (Namenode, RM, Hive, Oozie, Livy, Sqoop, HDFS, Spark...)

Event Bus

IaaS Layer

- Cluster A: Provisioning (kubernetes)
- Cluster B: Batch Job

Scalability
High Availability
MSA
Multi-Tenancy
Offering Data Analytics Service – Overall

- **DHP (Dynamic Hadoop Provisioning)**
  - Managed Hadoop Provisioning Service

- **Batch Pipeline**
  - Easy to Draw W/F by Drag and Drop

- **Real-time Pipeline**
  - Real-Time data processing

- **ML Modeler**
  - Spark ML Pipeline without Coding

- **R-Python Modeler**
  - R, Python Model Management

- **DL Modeler**
  - Deep Learning Life cycle management

- **Data Insight**
  - BI / Visualization Service

- **Cloud Search**
  - Provisioning ELK Stack
Dynamic Hadoop Provisioning (DHP)

- Managed Hadoop Provisioning Service
  - Optimized Hadoop / Spark Cluster in the containers
  - Provides various Hadoop Echoes
    - Hive / Oozie / Scoop / Livy
  - Caches end points into the shared cache for other services
  - Monitors container resources and Applications
Batch Pipeline (BP)

Easy to draw data pipeline by drag & drop components for Spark ETL

- Interactive mode
- Imports Spark ML Model for Operation
- Deploy each batch pipeline to one of clusters
- Monitors job status
- Collects MR/Spark log data from sand boxes
Real-Time Pipeline (RP)

Main Features

Drag & Drop GUI to draw Real-Time Pipeline

- Snapshot the streaming data
- Collect/Queue/Streaming
- Imports Spark ML model for operation
- Deploy each Real-Time Pipeline to one of clusters
- Supports Real-Time Charts
- Collects log data from sand boxes
ML Modeler

Main Features

- Spark ML Pipeline without Coding
- Distributed Spark ML Training
- Data and Model management
- SparkML Models can be operated by other services (BP/RP)
- SparkML Model -> Python Model to interact with python native libraries
**Main Features**

DL model Life cycle management

- Built-In algorithm mode for text and image classification by GUI
- Custom mode provides jupyter notebook for coding
- With the Horovod, DL Training can be Distributed across multiple GPUs
- Several trainings at the same time
- DL models deployed as API by Click

---

6 Simultaneous Training
Data Insight

Main Features

Data Visualization from many data sources

- Data Adaptors to collect data
- Data Set Management
- Drawing Charts
- Dynamic Dashboard with charts
- Supports Query
- Scheduler for updating Charts
Offering Public Cloud – Batch Predictions

Batch Predictions - Product Recommendation, Customer segmentations, Churn, Demand Prediction, Power Generation Prediction, Supply Chain Mgmt.

Combine services for Batch Predictions

Data Lake → ETL → DHP → Batch Pipeline → ML Modeler → Batch Pipeline → Data Insight

Yarn Containers → Spark Cluster (Sand Box) → DW
Offering Public – Real-Time Predictions

Real-Time Predictions - Intrusion detection, Detecting equipment failure, Real-Time Marketing, Fraud Detection, IOT services

Combine services for Real-Time Predictions

- DHP
- Batch Pipeline
- ML Modeler
- Real-time Pipeline
- Provisioning
- ETL
- Training
- Operation

Kafka ➔ ETL ➔ Yarn Containers ➔ ELK / DW

Cluster (Sand Box)
Offering On-Premise Cloud

Different aspects of delivering platform as On-Premise private cloud

- Deliver platform with combining services as customer requirements
  - BP+RP+Local Hadoop, DHP+BP, DHP+BP+RP+DI => PaaS is developed by MSA
- Approval System is required
  - Requesting Resources -> Admin Approval -> Provisioning the Resources
  - Separate GUI for administrators
- Dynamic Dashboard is provided
  - Each user can make a different dashboard according to the role
  - [Operators - Resources, jobs, ETL pipelines] [Analysts - Models, Data, Visualization]
- Customize platform and service as required
  - Add Specific Data processing as components on the service (BP/RP)
  - Customize Resource Size on the DHP (4core/16GB->5core/20GB)
- Integrating with Local Systems
  - Local Data Lake, Meta system, RDBMS, Kafka, ETL solutions, Job Scheduler, SSO etc
- Training users and operators
  - Training users for how to use the service
  - Training operators for how to manage the platform
Offering Hybrid Cloud

Combining On-Premise with Public Cloud

On-Premise Cloud
- Batch Pipeline
- ML Modeler
- Batch Pipeline
- DHP

AWS Public Cloud
- Docker Containers
- EKS
- Provisioning
- Cluster (Sand Box)
- EMR
- Provisioning
- Yarn Containers
- VMs

Data Lake

Reuse Analytics Model and Pipeline
Road Map

1st Step (2018)
AccuInsight+
Data Analytics Platform
for Industry Common
- DHP
- Batch Pipeline
- Realtime Pipeline
- BigQL
- ML Modeler
- DL Modeler
- Data Insight
- Cloud Search

2nd Step (2019)
[Industry AI Platform]
AccuInsight+
Diverse industries
- DHP
- Batch Pipeline
- Realtime Pipeline
- BigQL
- ML Modeler
- DL Modeler
- Data Insight
- Cloud Search
- R, Python
- Modeler
- Data Catalog

3rd Step (2020)
[Data Market Place]
- Seller
- Analysts
- Data Brokers
- Data Market Place
- Buyer
- Adver. Platform

- Upgrading platform for On-Premise Delivery
  - Some cases are used for AI Platform
- Developed More Services
  - R-Python Modeler, Data Catalog
- Hybrid Cloud with AWS Public Cloud
- AutoML

Connecting to Data Brokers and Data Market Places and Advert. Platforms
- Extending Hybrid cloud
- Adding Business APIs
Demo!!
End of Document