Data Pipeline & Data Lake

(in Google Cloud)

Megazone Google Cloud Team

Jung hoo Park

Agenda

- What is a Data Pipeline?
- Data Engineering in Google Cloud

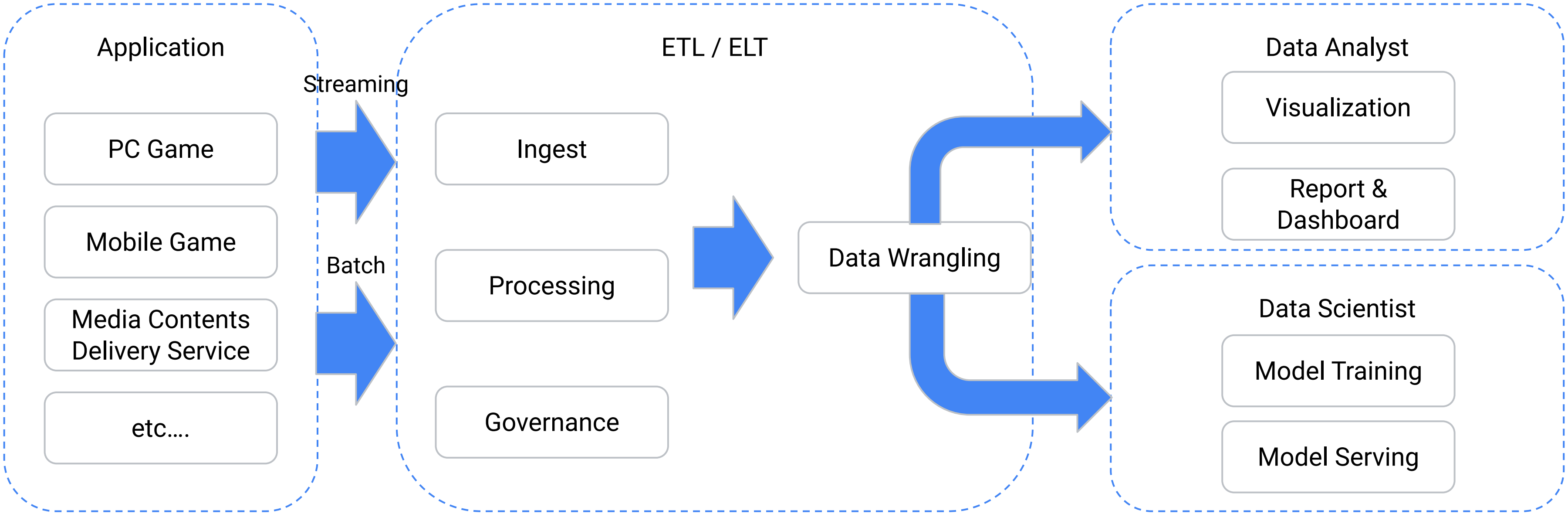
What is a Data Pipeline?

01

BIG
DATA
INFORMATION
STORAGE
LARGE
SOFTWARE
PETABYTES
ANALYSIS
RESEARCH
SEARCH
MANAGEMENT
TECHNOLOGIES
VOLUME
PROCESS
GOVERNMENT
SETS
MILLION
DEVELOPMENT
TECHNOLOGY
SCIENCE
FUTURE
NEW
BILLION
INTERNET
LEARNING
WORLD
BASED
SAN
NAS
EXABYTES
SIZE

Google
f
in
twitter

Conceptual Data Pipeline



Application

- PC Game
- Mobile Game
- Media Contents Delivery Service
- etc....

Streaming

Batch

ETL / ELT

- Ingest
- Processing
- Governance

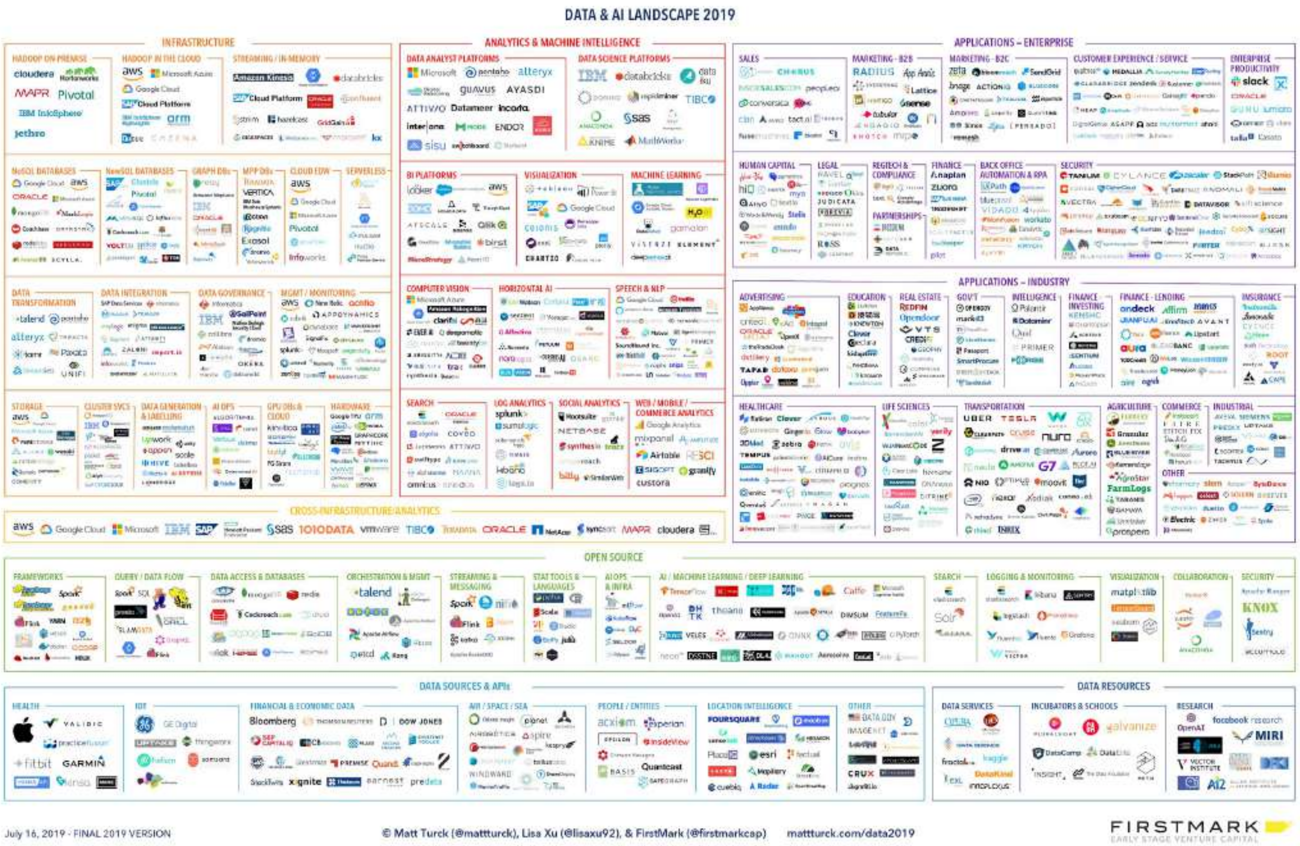
Data Wrangling

Data Analyst

- Visualization
- Report & Dashboard

Data Scientist

- Model Training
- Model Serving



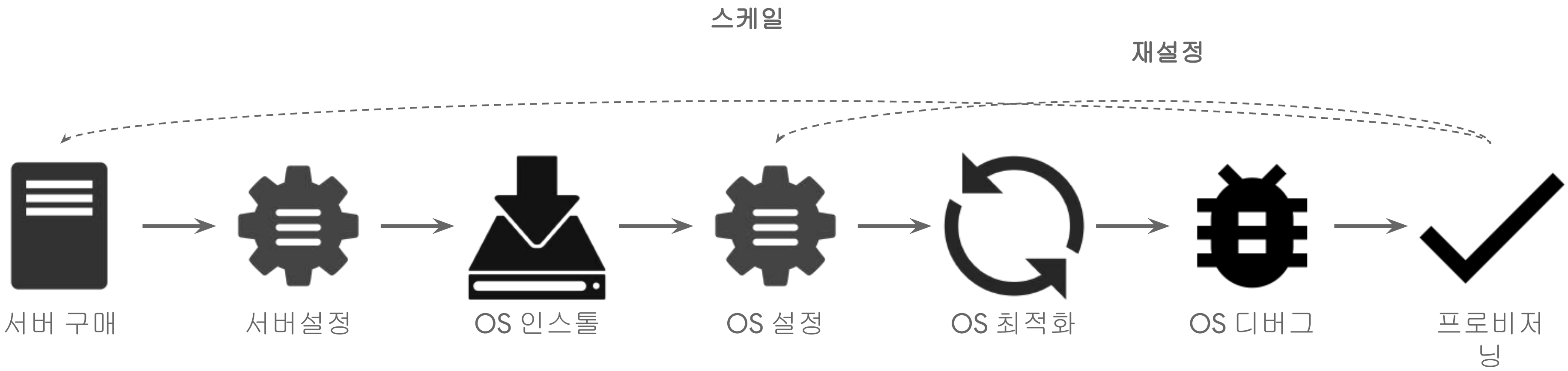
July 16, 2019 - FINAL 2019 VERSION

© Matt Turck (@mattturck), Lisa Xu (@lisaxu92), & FirstMark (@firstmarkcap) mattturck.com/data2019

Google Cloud

Cloud Innovator MEGAZONE

일반적인 데이터 처리 환경 프로비저닝



Google Cloud

Cloud Innovator MEGAZONE

Google Cloud

Google Cloud Service for Data Pipeline

02

Ingest

- Stackdriver Logging
- BigQuery Streaming API
- Cloud Pub/Sub
- Cloud IoT Core
- Transfer Appliance
- Transfer Service

Process

- Cloud Dataflow
- Cloud Dataproc
- Cloud Composer
- Cloud Dataprep
- Cloud Datalab

Store

- Cloud Storage
- BigQuery
- Cloud SQL
- Cloud Datastore
- Cloud Bigtable
- Cloud Spanner

Analyze

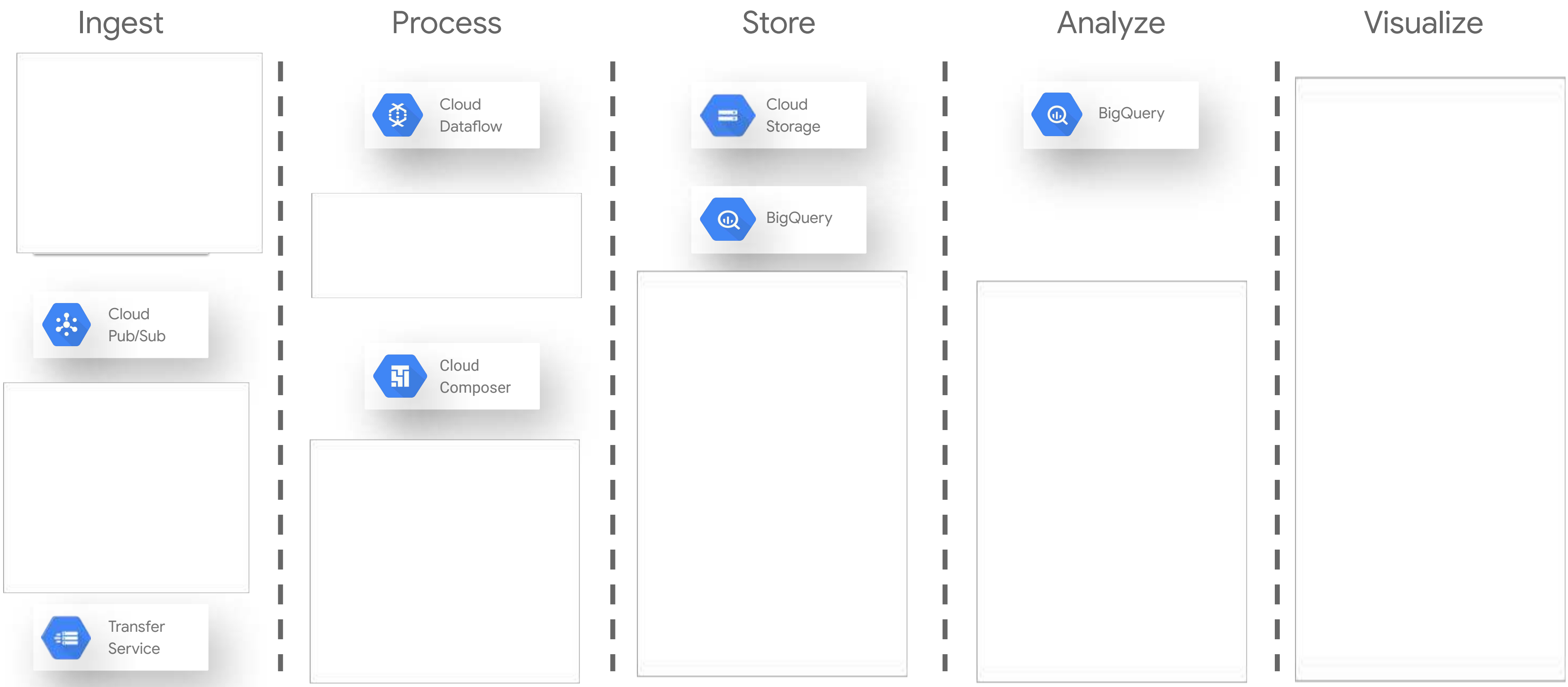
- BigQuery
- Cloud Dataprep
- Cloud Datalab

Visualize

- Data Studio
- Cloud Datalab
- Google Cloud Partner 3rd Party

Google Cloud

Cloud Innovator MEGAZONE



Ingest

- Cloud Pub/Sub
- Transfer Service

Process

- Cloud Dataflow
- Cloud Composer

Store

- Cloud Storage
- BigQuery

Analyze

- BigQuery

Visualize

Google Cloud

Cloud Innovator MEGAZONE

Ingest

Cloud
Pub/Sub

- Exactly-once processing
- Global by default
- No provisioning, auto-everything
- Seek and replay

Storage Transfer Service

- Transfer data from cloud to cloud
- Transfer data from bucket to bucket
- Centralized job management
- High-performance copies
- Data security

Processing

beam

Cloud Dataflow

- Auto-Scaling
- Streaming Engine
- Dataflow Shuffle
- Dataflow SQL
- Dataflow Template
- Inline Monitoring

Cloud Composer

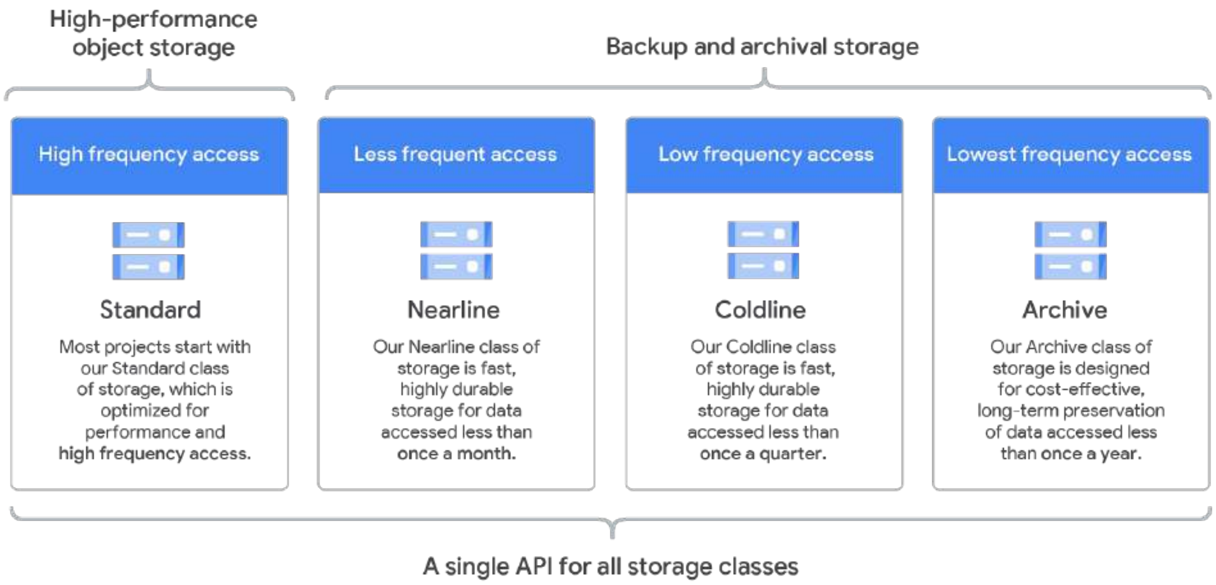
- Open Source
- Multi-Cloud
- Hybrid
- Integration
- Python Language
- Fully-Managed

Store & Analyze

Cloud Storage

- Object Life-cycle Management
- Object Version Management
- Retention Policies
- Pub/Sub Notifications for Cloud Storage
- Customer-managed encryption keys
- Cloud Audit Logs with Cloud Storage

Google Cloud Storage Class



High-performance object storage

Backup and archival storage

High frequency access

Standard

Most projects start with our Standard class of storage, which is optimized for performance and **high frequency access.**

Less frequent access

Nearline

Our Nearline class of storage is fast, highly durable storage for data accessed less than **once a month.**

Low frequency access

Coldline

Our Coldline class of storage is fast, highly durable storage for data accessed less than **once a quarter.**

Lowest frequency access

Archive

Our Archive class of storage is designed for cost-effective, long-term preservation of data accessed less **than once a year.**

A single API for all storage classes

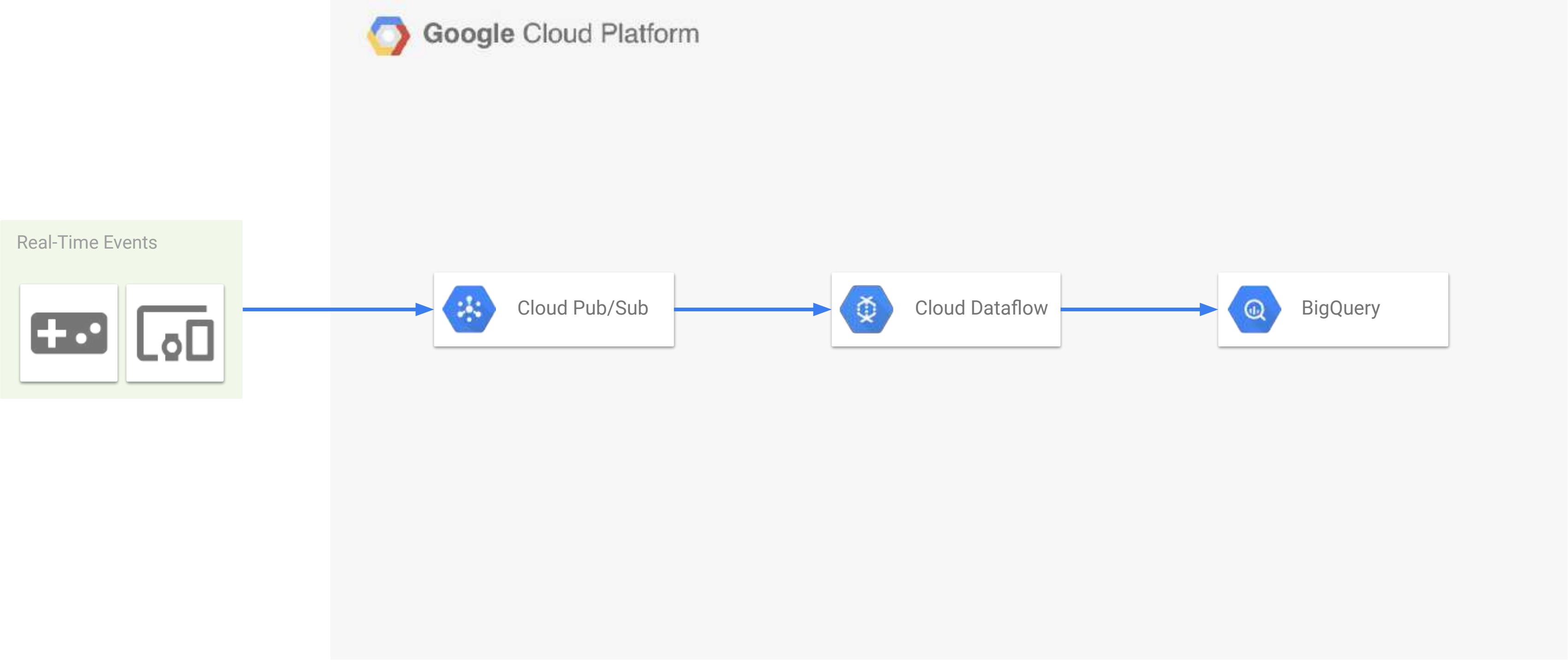
BigQuery

- Serverless
- Petabyte Scale
- Data Transfer Service
- Foundation for AI & BI
- Big data ecosystem integration
- Geo-expansion

Demo

03

Architecture: Demo Scenario



Thank you

Google Cloud

Cloud Innovator MEGAZONE