



Data Pipeline & Data Lake (in Google Cloud)

Megazone Google Cloud Team

Jung hoo Park

Google Cloud

MZ Cloud Innovator
MEGAZONE

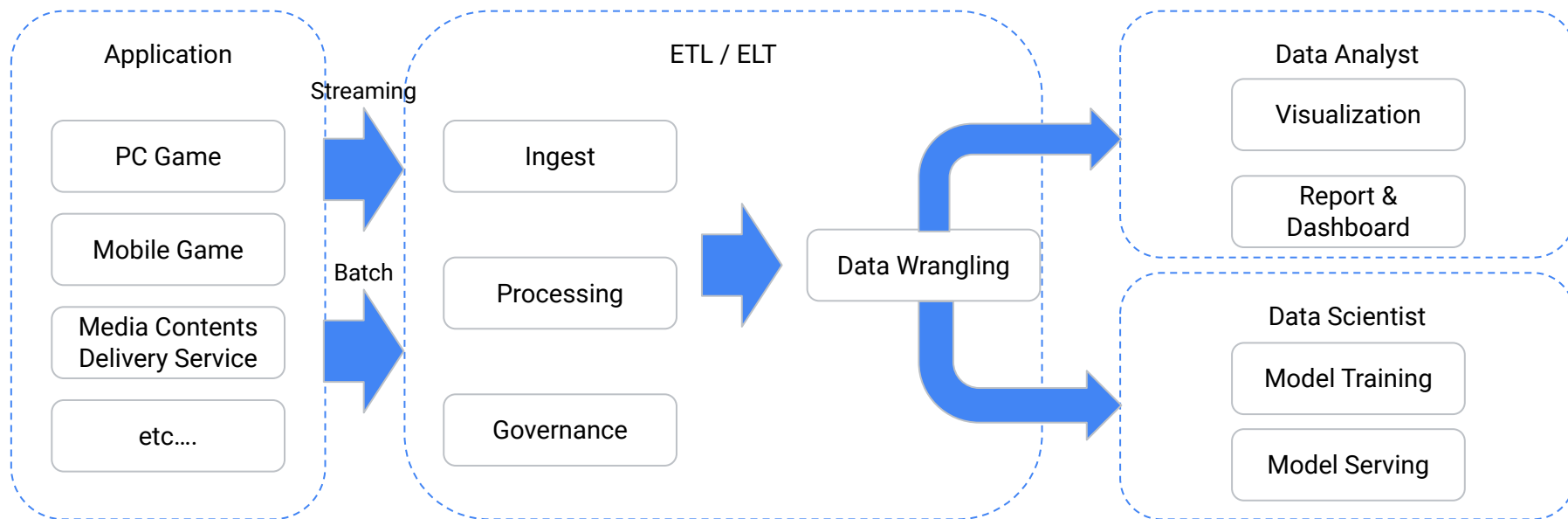
Agenda

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

What is a Data Pipeline?



Conceptual Data Pipeline



DATA & AI LANDSCAPE 2019

INFRASTRUCTURE			ANALYTICS & MACHINE INTELLIGENCE			APPLICATIONS - ENTERPRISE																	
HADOOP ON PREMISE cloudera, HDP MAAP Pivotal IBM InfoSphere Jettro	HADOOP IN THE CLOUD AWS, Microsoft Azure Google Cloud IBM Cloud Platform Oracle Cloud H2O	STREAMING / IN-MEMORY Amazon Kinesis, Azure Stream Analytics, Databricks Confluent, Redshift, Google Cloud, Databricks, Snowflake, Kudu, Hudi, Delta Lake	DATA ANALYST PLATFORMS Microsoft, Pentaho, Alteryx, Qlikview, Qlik Sense, Tableau, PowerBI, SAS, IBM, Alteryx, AVASDI, DataRobot, SAP, Oracle, Informatica, Tableau, SAS, IBM, Alteryx, ENDOR, SAP, Oracle, Informatica, Tableau, SAS, IBM, Alteryx, ENDOR	DATA SCIENCE PLATFORMS IBM, Microsoft Azure, Oracle, SAS, SAS Viya, SAS Analytics, SAS Predictive Analytics, SAS Visual Analytics, SAS Enterprise Miner, SAS Data Science Platform, SAS Decision Management, SAS Fraud Framework, SAS Risk Management, SAS Supply Chain, SAS Customer Experience, SAS Marketing, SAS Sales, SAS Retail, SAS Healthcare, SAS Financial Services, SAS Government	SALES CH2M, Microsoft Dynamics 365, Salesforce, SAP Sales Cloud, Oracle Sales Cloud, IBM Sales Hub, SAP Sales Hub, SAP Sales Hub Cloud Edition, SAP Sales Hub Cloud Edition, SAP Sales Hub Cloud Edition	MARKETING - B2C Adobe, SAP Marketing Cloud, Oracle Marketing Cloud, SAP Marketing Cloud, SAP Marketing Cloud, SAP Marketing Cloud, SAP Marketing Cloud, SAP Marketing Cloud, SAP Marketing Cloud	MARKETING - B2B SAP Marketing Cloud, SAP Marketing Cloud, SAP Marketing Cloud, SAP Marketing Cloud, SAP Marketing Cloud, SAP Marketing Cloud, SAP Marketing Cloud, SAP Marketing Cloud, SAP Marketing Cloud	CUSTOMER EXPERIENCE / SERVICE SAP Customer Experience, SAP Customer Experience, SAP Customer Experience, SAP Customer Experience, SAP Customer Experience, SAP Customer Experience, SAP Customer Experience, SAP Customer Experience	ENTERPRISE - PRODUCTIVITY Slack, Microsoft Office 365, SAP SuccessFactors, SAP SuccessFactors, SAP SuccessFactors, SAP SuccessFactors														
NO-SQL DATABASES Google Cloud, AWS, Oracle, SAP HANA, SAP HANA Cloud, SAP HANA Cloud, SAP HANA Cloud	NEW-SQL DATABASES Google Cloud, AWS, Oracle, SAP HANA, SAP HANA Cloud, SAP HANA Cloud, SAP HANA Cloud	GRAPH DBs Amazon Neptune, SAP HANA, SAP HANA Cloud, SAP HANA Cloud, SAP HANA Cloud	MPP DBs Amazon Redshift, SAP HANA, SAP HANA Cloud, SAP HANA Cloud, SAP HANA Cloud	CLOUD EDW Amazon Redshift, SAP HANA, SAP HANA Cloud, SAP HANA Cloud, SAP HANA Cloud	SERVERLESS Amazon Lambda, SAP HANA, SAP HANA Cloud, SAP HANA Cloud, SAP HANA Cloud	BI PLATFORMS SAP Business Intelligence, SAP Business Intelligence, SAP Business Intelligence, SAP Business Intelligence	VISUALIZATION SAP Analytics Cloud, SAP Analytics Cloud, SAP Analytics Cloud, SAP Analytics Cloud	MACHINE LEARNING SAP Machine Learning, SAP Machine Learning, SAP Machine Learning, SAP Machine Learning	LEGAL SAP Litigation, SAP Litigation, SAP Litigation, SAP Litigation	REGULATORY COMPLIANCE SAP Regulatory Compliance, SAP Regulatory Compliance, SAP Regulatory Compliance, SAP Regulatory Compliance	FINANCE SAP Finance, SAP Finance, SAP Finance, SAP Finance	BACK OFFICE AUTOMATION & RPA SAP Business Process Management, SAP Business Process Management, SAP Business Process Management, SAP Business Process Management	SECURITY SAP Security, SAP Security, SAP Security, SAP Security	OPERATIONS SAP Operations, SAP Operations, SAP Operations, SAP Operations									
CROSS-INFRASTRUCTURE-ANALYTICS			COMPUTER VISION			HORIZONTAL AI			SPEECH & NLP														
DATA TRANSFORMATION SAP Data Intelligence, SAP Data Intelligence, SAP Data Intelligence, SAP Data Intelligence	DATA INTEGRATION SAP Data Integration, SAP Data Integration, SAP Data Integration, SAP Data Integration	DATA GOVERNANCE SAP Data Governance, SAP Data Governance, SAP Data Governance, SAP Data Governance	BI/MAT. MONITORING SAP Business Intelligence, SAP Business Intelligence, SAP Business Intelligence, SAP Business Intelligence	COMPUTER VISION SAP Computer Vision, SAP Computer Vision, SAP Computer Vision, SAP Computer Vision	HORIZONTAL AI SAP Horizontal AI, SAP Horizontal AI, SAP Horizontal AI, SAP Horizontal AI	SPEECH & NLP SAP Speech & NLP, SAP Speech & NLP, SAP Speech & NLP, SAP Speech & NLP	HEALTHCARE SAP Healthcare, SAP Healthcare, SAP Healthcare, SAP Healthcare	REAL ESTATE SAP Real Estate, SAP Real Estate, SAP Real Estate, SAP Real Estate	GOVT. & PUBLIC SAP Government, SAP Government, SAP Government, SAP Government	FINANCE - INVESTING SAP Finance Investing, SAP Finance Investing, SAP Finance Investing, SAP Finance Investing	FINANCE - LENDING SAP Finance Lending, SAP Finance Lending, SAP Finance Lending, SAP Finance Lending	INSURANCE SAP Insurance, SAP Insurance, SAP Insurance, SAP Insurance	AGRICULTURE SAP Agriculture, SAP Agriculture, SAP Agriculture, SAP Agriculture	COMMERCE SAP Commerce, SAP Commerce, SAP Commerce, SAP Commerce	INDUSTRIAL SAP Industrial, SAP Industrial, SAP Industrial, SAP Industrial								
OPEN SOURCE			DATA SOURCES & API			AI / MACHINE LEARNING / DEEP LEARNING			SEARCH			LOGGING & MONITORING			VISUALIZATION			COLLABORATION			SECURITY		
FRAMEWORKS TensorFlow, PyTorch, Keras, MXNet, Caffe, Theano, TensorFlow, PyTorch, Keras, MXNet, Caffe, Theano	QUERY / DATA FLOW Apache Spark, Apache Flink, Apache Airflow, Apache Beam, Apache Storm, Apache Kudu, Apache Hudi, Apache Delta Lake	DATA ACCESS & DATABASES SAP Data Access, SAP Data Access, SAP Data Access, SAP Data Access	ORCHESTRATION & MGMT. SAP Orchestration, SAP Orchestration, SAP Orchestration, SAP Orchestration	STREAMING & MESSAGING SAP Streaming, SAP Streaming, SAP Streaming, SAP Streaming	DATA SOURCES & API SAP Data Sources & API, SAP Data Sources & API, SAP Data Sources & API, SAP Data Sources & API	AI / MACHINE LEARNING / DEEP LEARNING SAP AI, SAP AI, SAP AI, SAP AI	SEARCH SAP Search, SAP Search, SAP Search, SAP Search	LOGGING & MONITORING SAP Logging & Monitoring, SAP Logging & Monitoring, SAP Logging & Monitoring, SAP Logging & Monitoring	VISUALIZATION SAP Visualization, SAP Visualization, SAP Visualization, SAP Visualization	COLLABORATION SAP Collaboration, SAP Collaboration, SAP Collaboration, SAP Collaboration	SECURITY SAP Security, SAP Security, SAP Security, SAP Security	HEALTH SAP Health, SAP Health, SAP Health, SAP Health	FINANCIAL & ECONOMIC DATA SAP Financial & Economic Data, SAP Financial & Economic Data, SAP Financial & Economic Data, SAP Financial & Economic Data	PEOPLE / ENTERPRISE SAP People / Enterprise, SAP People / Enterprise, SAP People / Enterprise, SAP People / Enterprise	EDUCATION INTELLIGENCE SAP Education Intelligence, SAP Education Intelligence, SAP Education Intelligence, SAP Education Intelligence	OTHER DATA SAP Other Data, SAP Other Data, SAP Other Data, SAP Other Data	DATA SERVICES SAP Data Services, SAP Data Services, SAP Data Services, SAP Data Services	INCUBATORS & SCHOOLS SAP Incubators & Schools, SAP Incubators & Schools, SAP Incubators & Schools, SAP Incubators & Schools	RESEARCH SAP Research, SAP Research, SAP Research, SAP Research				

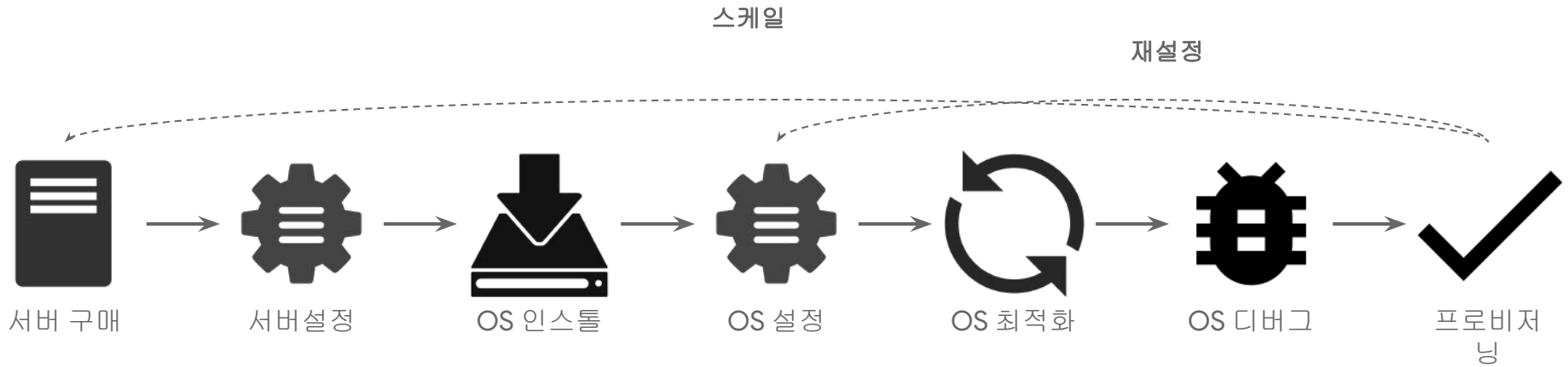
July 16, 2019 - FINAL 2019 VERSION

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

FIRSTMARK
EARLY STAGE VENTURE CAPITAL



일반적인 데이터 처리 환경 프로비저닝





Google Cloud

Google Cloud Service for Data Pipeline



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

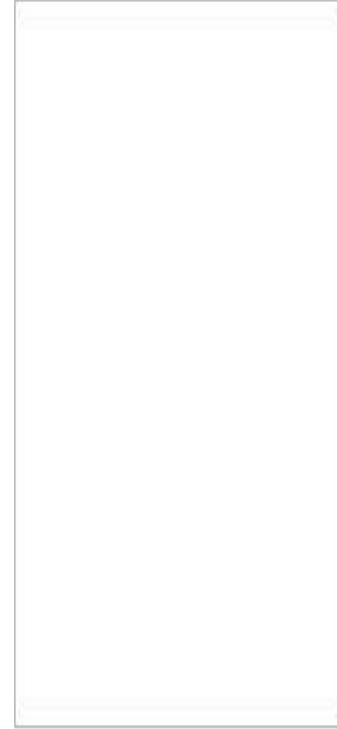
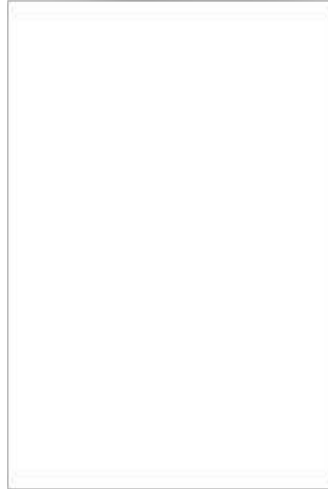
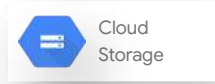
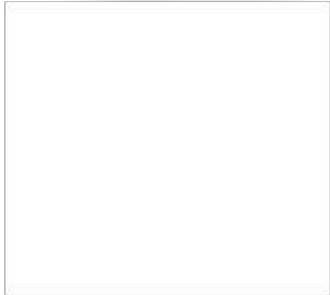
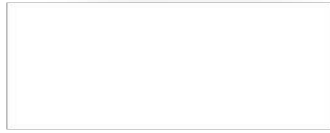
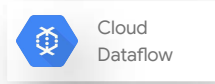
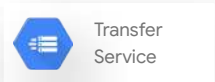
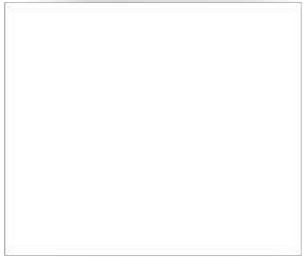
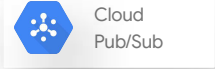
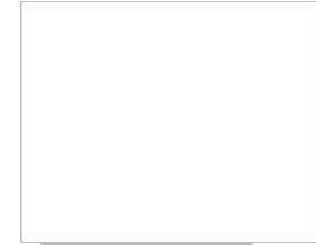
Ingest

Process

Store

Analyze

Visualize



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





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

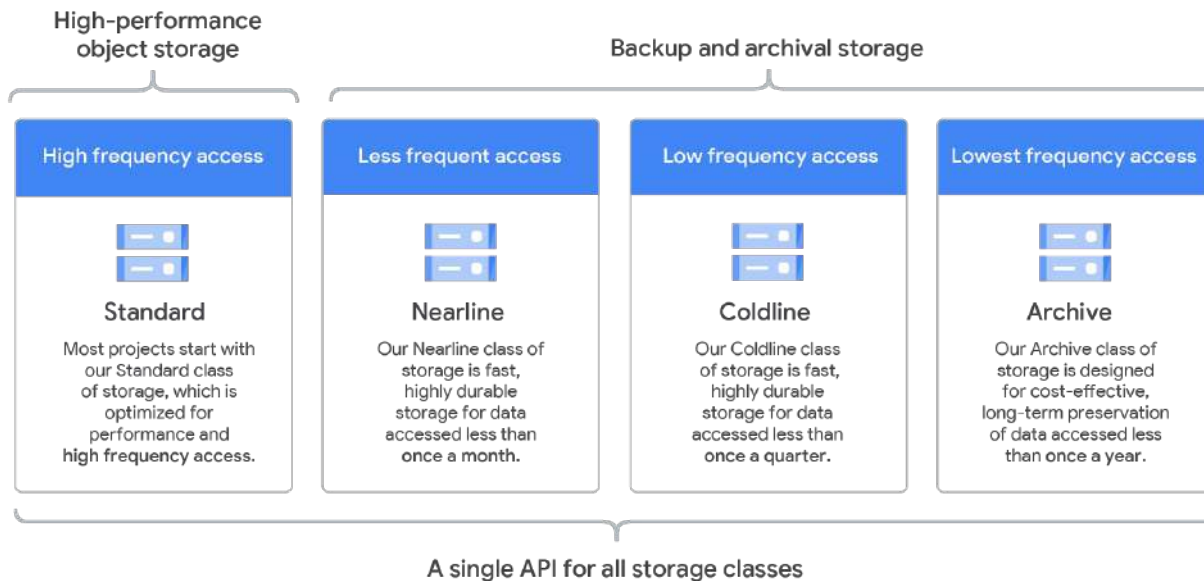




**Cloud
Storage**

- **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





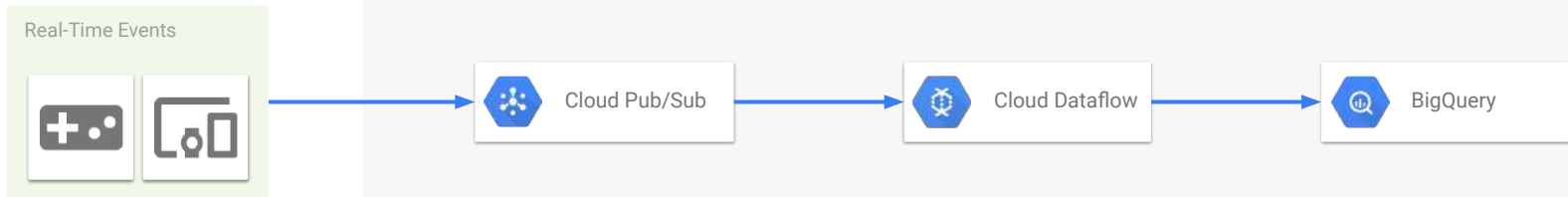
BigQuery

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

Demo



 Google Cloud Platform





Thank you