Overview - Analytics - Cloudera Data Offload (Data Lake Offload)

Work Tasks	Details	Primary Responsibilities
Scope Brief	Modernization on existing Hadoop / Cloudera platform: Offload cold data storage platform to GCP over a period of time Migrate HBase data to Cloud Bigtable for Call Detail Record purpose Offload existing Data Monetization, Deploy sandbox platform for external partners environment *Excludes new analytics use-cases previously proposed: (i) Kafka streaming job conversion to Dataflow.	• Assessment of Use Cases Planning • Review requirement, plan and overview documents. • Create offloading plan by use cases scenarios • Communicate initial offloading scheduling to app teams • Data Storage Offload - Data movement assessment, planning, and design • Cold Data Storage Offload - Planning data locations with permissions in mind • Cold Data Storage Offload - Planning an incremental data movement • Cold Data Storage Offload - Keeping your data synchronized in a hybrid solution • Cold Data Storage Offload - Configuring access to your data • Cold Data Storage Offload - Validating data transfers • Migrate HBase to Cloud Bigtable for (CDR) - Plan the migration: collect details from HBase • Migrate HBase to Cloud Bigtable for (CDR) - Export the HBase table to Cloud Storage • Migrate HBase to Cloud Bigtable for (CDR) - Migrate the sequence files from HDFS to Cloud Storage • Migrating from HBase to Cloud Bigtable for (CDR) - Verify the imported data inside Bigtable • Data Monetization - Provide sandbox for Partners/customers where they have their own compute and storage in Google Cloud. Organized as a Google Cloud projects depending on the requirements.
in-scope Dimensions	**Total Cold Data Lake Workload (Existing in data lake is Approx. 5.9 PB with 2.5 replication factor [5.9PB / 2.5 = Approx. 2.36 PB])** • Y1 (5.9 PB with 2.5 replication factor) ~ 2.36 PB • Y2 (29% YoY growth) ~ 3.04 PB • Y3 (17% YoY growth) ~ 3.56 PB • Y4 (12% YoY growth) ~ 3.99 PB • Y5 (15% YoY growth) ~ 4.59 PB	**Total HBase Workload (Approx. 640 TB)** • Y1 (BigTable 640 TB HDD Storage) • Y2 (28% YoY growth) ~ 689 TB • Y3 (5.2% YoY growth) ~ 725 TB • Y4 (6% growth) ~ 769 TB • Y5 (7% YoY growth) ~ 822 TB	**Total CVM Workload (TB)** • Y2 (Based on existing compute 40% of Impala cluster) ~ 600 TB • Y3 (35% YoY data growth, assuming same as data lake growth) ~ 702 TB • Y4 (12% data growth, assuming same as data lake growth) ~ 786 TB • Y5 (15% YoY data growth (assuming same as data lake growth) ~ 904 TB
Deliverables	• Cold Data Storage Offload (Kickoff document, Assessment Plan document, Offloading Design document, Implementation Report document, UAT document) • Migrate HBase data to Cloud Bigtable for Call Detail Record (Implementation document, UAT document) • Offload existing Data Monetization (Implementation document, UAT document) • Deploy sandbox platform for external partners environment (Implementation document, UAT document)		
Skillset Expertise	• GCP Certified Cloud Engineers (Professional Cloud Solution Architects & Professional DevOps Engineers) • GCP Certified Data Engineers (Professional Data Engineers as Data Engineer, Data Scientist, Data Visualization Engineer, and Business Analyst) • Certified Project Manager • Competencies and Certified Resources for On-Premise Stack (Cloudera, Greenplum, Kafka, Tableau, etc)		

Overview - Analytics - Data Monetization (Existing Use - Cases)

Work Tasks	Details	Primary Responsibilities	
Scope Brief	• Modernization on existing Hadoop / Cloudera platform: • Offload cold data storage platform to GCP over a period of time • Migrate HBase data to Cloud Bigtable for Call Detail Record purpose • Offload existing Data Monetization, • Deploy sandbox platform for external partners environment • *Excludes new analytics use-cases previously proposed: (i) Kafka streaming job conversion to Dataflow.	• Assessment of Use Cases Planning • Review requirement, plan and overview documents. • Create offloading plan by use cases scenarios • Communicate initial offloading scheduling to app teams • Data Storage Offload - Data movement assessment, planning, and design • Cold Data Storage Offload - Planning data locations with permissions in mind • Cold Data Storage Offload - Planning an incremental data movement • Cold Data Storage Offload - Keeping your data synchronized in a hybrid solution • Cold Data Storage Offload - Configuring access to your data • Cold Data Storage Offload - Validating data transfers • Migrate HBase to Cloud Bigtable for (CDR) - Plan the migration: collect details from HBase • Migrate HBase to Cloud Bigtable for (CDR) - Export the HBase table to Cloud Storage • Migrate HBase to Cloud Bigtable for (CDR) - Migrate the sequence files from HDFS to Cloud Storage • Migrating from HBase to Cloud Bigtable for (CDR) - Verify the imported data inside Bigtable • Data Monetization - Provide sandbox for Partners/customers where they have their own compute and storage in Google Cloud. Organized as a Google Cloud projects depending on the requirements.	
in-scope Dimensions	**Total BigQuery Workload (TB)** • Y1 (Based on input from Indosat, estimation using existing compute 15% of data lake cluster 49 servers) ~ 130 TB • Y2 (29% YoY growth (same as data lake growth) ~ 168 TB • Y3 (17% YoY data growth (same as data lake growth) ~ 196 TB • Y4 (12% data growth (same as data lake growth) ~ 220 TB • Y5 (15% YoY growth (same as data lake growth) ~ 253 TB	**Total Cloud Storage Workload (GB)** • Y1 (30% of BigQuery active storage) ~ 39 GB • Y2 (29% YoY growth) ~ 50 GB • Y3 (17% YoY growth) ~ 59 GB • Y4 (12% YoY growth) ~ 66 GB • Y5 (15% YoY growth) ~ 76 GB	**Total DLP Transformation Workload** • 1 TB data per day, so 30 TB per month. Tokenization will cover approximately 5% of total data volume. • 1 TB data per day, so 30 TB per month. Tokenization will cover approximately 5% of total data volume. • 1 TB data per day, so 30 TB per month. Tokenization will cover approximately 5% of total data volume. • 1 TB data per day, so 30 TB per month. Tokenization will cover approximately 5% of total data volume. • 1 TB data per day, so 30 TB per month. Tokenization will cover approximately 5% of total data volume.~ 904 TB
Deliverables	• Cold Data Storage Offload (Kickoff document, Assessment Plan document, Offloading Design document, Implementation Report document, UAT document) • Migrate HBase data to Cloud Bigtable for Call Detail Record (Implementation document, UAT document) • Offload existing Data Monetization (Implementation document, UAT document) • Deploy sandbox platform for external partners environment (Implementation document, UAT document)		
Skillset Expertise	• GCP Certified Cloud Engineers (Professional Cloud Solution Architects & Professional DevOps Engineers) • GCP Certified Data Engineers (Professional Data Engineers as Data Engineer, Data Scientist, Data Visualization Engineer, and Business Analyst) • Certified Project Manager • Competencies and Certified Resources for On-Premise Stack (Cloudera, Greenplum, Kafka, Tableau, etc)		

Overview - Analytics - Additional Capabilities

Work Tasks	Details	Primary Responsibilities
Scope Brief	• Reporting and visualization for B2B Indosat Phone Traffic & Revenue • Data Governance with Collibra	• Assessment of Use Cases Planning • Review requirement, plan and overview documents. • Create offloading plan by use cases scenarios • Define application dependencies • Identify validated source items • Identify technical constraints • Communicate initial offloading scheduling to app teams • Reporting and visualization discovery and kickoff • Create an end-to-end setup that enables a BI tool to use data from a Hadoop/BigQUery environment. • Authenticate and authorize user requests. • Set up and use secure communication channels between the BI tool and the cluster. • Load data for reporting to BigQuery • Create reporting and visualization for B2B IO Phone Traffic & Revenue in Looker
in-scope Dimensions	**Total BigQuery Offload** • Y1 (10% of data lake cluster (32 servers). (total 700 slots). 400 slots for reporting) ~ 365 GB • Y2 (13.6% YoY compute growth. (total 800 slots) ~ 730 GB • Y3 (9% YoY compute growth. (total 900 slots) ~ 1095 GB • Y4 (7% YoY compute growth. (total 1000 slots) ~ 1460 GB • Y5 (7% YoY compute growth. This slots will be shared by cold data queries and reporting (total 1000 slots). ~ 1825 GB	**Total Looker Workload** • Year-1 Manage 2 Admin Users, 10 User Editors, 40 View users (manage cold data queries and reporting (400 slots for reporting.) • Year-2 Manage 2 Admin Users, 10 User Editors, 40 View users (total 800 slots). 13.6% YoY compute growth.) • Year-3 Manage 2 Admin Users, 10 User Editors, 40 View users (total 900 slots). 9% YoY compute growth.) • Year-4 Manage 2 Admin Users, 10 User Editors, 40 View users (total 1000 slots). 7% YoY compute growth.) • Year-5 Manage 2 Admin Users, 10 User Editors, 40 View users (total 1100 slots). 9% YoY compute growth.)
Deliverables	• Assessment Plan document • BI Offloading Design document • Implementation Report document • Validation document	
Skillset Expertise	• GCP Certified Cloud Engineers (Professional Cloud Solution Architects & Professional DevOps Engineers) • GCP Certified Data Engineers (Professional Data Engineers as Data Engineer, Data Scientist, Data Visualization Engineer, and Business Analyst) • Certified Project Manager • Competencies and Certified Resources for On-Premise Stack (Cloudera, Greenplum, Kafka, Tableau, etc)	

Overview - Analytics - New Use Cases Implementation (out-of-Scope)

Work Tasks	Details	Main responsibilities
Scope Brief	Offloading streaming data from Kafka. This streaming data needs to be consumed by 3rd parties.	• Kafka UC Stream - Load from geolocation data through Kafka to BigQuery • Processed from Kafka topic, simple transformation, insert into DB • Create and organize tokenize PII. Existing tokenization using Spark Streaming • Configure additional Tokenization needs • Transport streaming instead of storing in data lake • Managing the related use cases data monetization.new use cases • AI/ML Platform - to build machine learning models, starting from experimentation, training, and productionizing. • Partner/customer can access the data in multiple ways: query, API, or a portal if needed. This will depend on the requirements. • Partner/customer can have a sandbox, where they have their own compute and storage in Google Cloud. This sandbox is organized as a Google Cloud projects, so IO can organize depending on the requirements.
in-scope Dimensions	**Pub/Sub Workload(s)** • Streaming use case start in Y2 • Y2 (Initial 5.2 TB daily. Kafka YoY data growth 38%) ~ 215 TB • Y3 (Kafka data growth 9%) ~ 235 TB • Y4 (Kafka data growth 9.7%) ~ 257 TB • Y5 (Kafka YoY data growth 11%) ~ 286 TB	**Dataflow Streaming (Workload)s** • Streaming use case start in Y2 • Y2 (Disk 10% of existing servers (existing 48 TB). Dataflow 3 nodes, each 480 GB) ~ 1440 GB • Y3 (Assuming 4 nodes, each 300 GB) ~ 1200 GB • Y4 (Assuming 6 nodes, each 300 GB) ~ 1800 GB • Y5 (Assuming 8 nodes, each 300 GB) ~ 2400 GB
Deliverables	• Assessment Plan Document • Offloading Streaming Design Document • Implementation Report Document • UAT Document	
Resources Utilized	• GCP Certified Cloud Engineers (Professional Cloud Solution Architects) • GCP Certified Data Engineers (Professional Data Engineers as Data Engineer, Data Scientist, Data Visualization Engineer, and Business Analyst) • Certified Project Manager