**Case Study: 1 – MJTelco**

**Company Overview**

**MJTelco is a startup that plans to build networks in rapidly growing, underserved markets around the world. The company has patents for innovative optical communications hardware. Based on these patents, they can create many reliable, high-speed backbone links with inexpensive hardware.  
Company Background  
Founded by experienced telecom executives, MJTelco uses technologies originally developed to overcome communications challenges in space. Fundamental to their operation, they need to create a distributed data infrastructure that drives real-time analysis and incorporates machine learning to continuously optimize their topologies. Because their hardware is inexpensive, they plan to overdeploy the network allowing them to account for the impact of dynamic regional politics on location availability and cost. Their management and operations teams are situated all around the globe creating many-to- many relationship between data consumers and provides in their system. After careful consideration, they decided public cloud is the perfect environment to support their needs.  
Solution Concept  
MJTelco is running a successful proof-of-concept (PoC) project in its labs. They have two primary needs:  
Scale and harden their PoC to support significantly more data flows generated when they ramp to more than 50,000 installations.**

**Refine their machine-learning cycles to verify and improve the dynamic models they use to control topology definition.  
MJTelco will also use three separate operating environments ?development/test, staging, and production ?  
to meet the needs of running experiments, deploying new features, and serving production customers.  
Business Requirements  
Scale up their production environment with minimal cost, instantiating resources when and where needed in an unpredictable, distributed telecom user community. Ensure security of their proprietary data to protect their leading-edge machine learning and analysis.  
Provide reliable and timely access to data for analysis from distributed research workers Maintain isolated environments that support rapid iteration of their machine-learning models without affecting their customers.  
Technical Requirements  
Ensure secure and efficient transport and storage of telemetry data Rapidly scale instances to support between 10,000 and 100,000 data providers with multiple flows each.  
Allow analysis and presentation against data tables tracking up to 2 years of data storing approximately  
100m records/day  
Support rapid iteration of monitoring infrastructure focused on awareness of data pipeline problems both in telemetry flows and in production learning cycles.  
CEO Statement  
Our business model relies on our patents, analytics and dynamic machine learning. Our inexpensive hardware is organized to be highly reliable, which gives us cost advantages. We need to quickly stabilize our large distributed data pipelines to meet our reliability and capacity commitments.  
CTO Statement  
Our public cloud services must operate as advertised. We need resources that scale and keep our data secure. We also need environments in which our data scientists can carefully study and quickly adapt our models. Because we rely on automation to process our data, we also need our development and test environments to work as we iterate.  
CFO Statement  
The project is too large for us to maintain the hardware and software required for the data and analysis.  
Also, we cannot afford to staff an operations team to monitor so many data feeds, so we will rely on automation and infrastructure. Google Cloud's machine learning will allow our quantitative researchers to work on our high-value problems instead of problems with our data pipelines.**

**You create a new report for your large team in Google Data Studio 360. The report uses Google BigQuery as its data source. It is company policy to ensure employees can view only the data associated with their region, so you create and populate a table for each region. You need to enforce the regional access policy to the data.  
Which two actions should you take? (Choose two.)**

**A.** Adjust the settings for each view to allow a related region-based security group view access.

**B.** Ensure all the tables are included in global dataset.

**C.** Adjust the settings for each table to allow a related region-based security group view access.

**D.** Ensure each table is included in a dataset for a region.

**E.** Adjust the settings for each dataset to allow a related region-based security group view access.

Ans: A, D

**MJTelco needs you to create a schema in Google Bigtable that will allow for the historical analysis of the last 2 years of records. Each record that comes in is sent every 15 minutes, and contains a unique identifier of the device and a data record. The most common query is for all the data for a given device for a given day. Which schema should you use?**

**A.** Rowkey: device\_idColumn data: date, data\_point

**B.** Rowkey: date#data\_pointColumn data: device\_id

**C.** Rowkey: date#device\_idColumn data: data\_point

**D.** Rowkey: data\_pointColumn data: device\_id, date

**E.** Rowkey: dateColumn data: device\_id, data\_point

Ans: D

**Given the record streams MJTelco is interested in ingesting per day, they are concerned about the cost of Google BigQuery increasing. MJTelco asks you to provide a design solution. They require a single large data table called tracking\_table. Additionally, they want to minimize the cost of daily queries while performing fine-grained analysis of each day's events. They also want to use streaming ingestion. What should you do?**

**A.** Create sharded tables for each day following the pattern tracking\_table\_YYYYMMDD.

**B.** Create a table called tracking\_table and include a DATE column.

**C.** Create a partitioned table called tracking\_table and include a TIMESTAMP column.

**D.** Create a table called tracking\_table with a TIMESTAMP column to represent the day.

Ans: C

**You need to compose visualization for operations teams with the following requirements:  
\* Telemetry must include data from all 50,000 installations for the most recent 6 weeks (sampling once every minute)  
\* The report must not be more than 3 hours delayed from live data.  
\* The actionable report should only show suboptimal links.  
\* Most suboptimal links should be sorted to the top.  
\* Suboptimal links can be grouped and filtered by regional geography.  
\* User response time to load the report must be <5 seconds.  
You create a data source to store the last 6 weeks of data, and create visualizations that allow viewers to see multiple date ranges, distinct geographic regions, and unique installation types. You always show the latest data without any changes to your visualizations. You want to avoid creating and updating new visualizations each month. What should you do?**

**A.** Load the data into relational database tables, write a Google App Engine application that queries all rows, summarizes the data across each criteria, and then renders results using the Google Charts and visualization API.

**B.** Export the data to a spreadsheet, compose a series of charts and tables, one for each possible combination of criteria, and spread them across multiple tabs.

**C.** Look through the current data and compose a series of charts and tables, one for each possible combination of criteria.

**D.** Look through the current data and compose a small set of generalized charts and tables bound to criteria filters that allow value selection.

Ans: D

**Case Study – 2 - Flowlogistic**

**Company Overview**

**Flowlogistic is a leading logistics and supply chain provider. They help businesses throughout the world manage their resources and transport them to their final destination. The company has grown rapidly, expanding their offerings to include rail, truck, aircraft, and oceanic shipping.  
Company Background  
The company started as a regional trucking company, and then expanded into other logistics market.  
Because they have not updated their infrastructure, managing and tracking orders and shipments has become a bottleneck. To improve operations, Flowlogistic developed proprietary technology for tracking shipments in real time at the parcel level. However, they are unable to deploy it because their technology stack, based on Apache Kafka, cannot support the processing volume. In addition, Flowlogistic wants to further analyze their orders and shipments to determine how best to deploy their resources.  
Solution Concept  
Flowlogistic wants to implement two concepts using the cloud:  
Use their proprietary technology in a real-time inventory-tracking system that indicates the location of  
their loads Perform analytics on all their orders and shipment logs, which contain both structured and unstructured data, to determine how best to deploy resources, which markets to expand info. They also want to use predictive analytics to learn earlier when a shipment will be delayed.**

**Existing Technical Environment**

**Flowlogistic architecture resides in a single data center:**

**Databases  
8 physical servers in 2 clusters  
- SQL Server - user data, inventory, static data  
3 physical servers  
- Cassandra - metadata, tracking messages  
10 Kafka servers - tracking message aggregation and batch insert  
Application servers - customer front end, middleware for order/customs** **60 virtual machines across 20 physical servers  
- Tomcat - Java services  
- Nginx - static content  
- Batch servers  
Storage appliances** **- iSCSI for virtual machine (VM) hosts  
- Fibre Channel storage area network (FC SAN) - SQL server storage  
- Network-attached storage (NAS) image storage, logs, backups  
Apache Hadoop /Spark servers** **- Core Data Lake  
- Data analysis workloads  
20 miscellaneous servers** **- Jenkins, monitoring, bastion hosts,  
Business Requirements  
Build a reliable and reproducible environment with scaled panty of production.** **Aggregate data in a centralized Data Lake for analysis** **Use historical data to perform predictive analytics on future shipments** **Accurately track every shipment worldwide using proprietary technology** **Improve business agility and speed of innovation through rapid provisioning of new resources** **Analyze and optimize architecture for performance in the cloud** **Migrate fully to the cloud if all other requirements are met** **Technical Requirements  
Handle both streaming and batch data** **Migrate existing Hadoop workloads** **Ensure architecture is scalable and elastic to meet the changing demands of the company.** **Use managed services whenever possible** **Encrypt data flight and at rest** **Connect a VPN between the production data center and cloud environment** **SEO Statement  
We have grown so quickly that our inability to upgrade our infrastructure is really hampering further growth and efficiency. We are efficient at moving shipments around the world, but we are inefficient at moving data around.  
We need to organize our information so we can more easily understand where our customers are and what they are shipping.**

**CTO Statement  
IT has never been a priority for us, so as our data has grown, we have not invested enough in our technology. I have a good staff to manage IT, but they are so busy managing our infrastructure that I cannot get them to do the things that really matter, such as organizing our data, building the analytics, and figuring out how to implement the CFO' s tracking technology.**

**CFO Statement  
Part of our competitive advantage is that we penalize ourselves for late shipments and deliveries. Knowing where out shipments are at all times has a direct correlation to our bottom line and profitability.  
Additionally, I don't want to commit capital to building out a server environment.**

**Flowlogistic wants to use Google BigQuery as their primary analysis system, but they still have Apache Hadoop and Spark workloads that they cannot move to BigQuery. Flowlogistic does not know how to store the data that is common to both workloads. What should they do?**

**A.** Store he common data in the HDFS storage for a Google Cloud Dataproc cluster.

**B.** Store the common data encoded as Avro in Google Cloud Storage.

**C.** Store the common data in BigQuery as partitioned tables.

**D.** Store the common data in BigQuery and expose authorized views.

Ans: D

**Flowlogistic's CEO wants to gain rapid insight into their customer base so his sales team can be better  
informed in the field. This team is not very technical, so they've purchased a visualization tool to simplify  
the creation of BigQuery reports. However, they've been overwhelmed by all the data in the table, and are  
spending a lot of money on queries trying to find the data they need. You want to solve their problem in the  
most cost-effective way. What should you do?**

**A.** Export the data into a Google Sheet for virtualization.

**B.** Create an additional table with only the necessary columns.

**C.** Create a view on the table to present to the virtualization tool.

**D.** Create identity and access management (IAM) roles on the appropriate columns, so only they appear  
in a query.

Ans: C

**Flowlogistic is rolling out their real-time inventory tracking system. The tracking devices will all send package-tracking messages, which will now go to a single Google Cloud Pub/Sub topic instead of the Apache Kafka cluster. A subscriber application will then process the messages for real-time reporting and store them in Google BigQuery for historical analysis. You want to ensure the package data can be analyzed over time.  
Which approach should you take?**

**A.** Attach the timestamp on each message in the Cloud Pub/Sub subscriber application as they are received.

**B.** Attach the timestamp and Package ID on the outbound message from each publisher device as they are sent to Cloud Pub/Sub.

**C.** Use the NOW () function in BigQuery to record the event's time.

**D.** Use the automatically generated timestamp from Cloud Pub/Sub to order the data.

Ans: B