Big Data Solution for Healthcare

Anusha Bhat
31-013
October 13, 2024

Agenda

Data Collection

Data Analytics

Big Data Technology Solution

Data and Capacity Sizing

Total Monthly Cost

Reference Architecture Diagram

Data Collection

- Need to store data regarding patient vitals, demographic information, social determinants of health (SDOH), and past medical history; including but not limited to:
  - Age, height, weight
  - Blood pressure, heart rate, oxygen saturation levels, temperature
  - Education level, race, financial/economic stability, living conditions/environment
  - Family or personal history of any disease, smoking history, mental health wellness check
- An Electronic Health Record (EHRs) copy should also be stored for each patient
  - Includes data on patient's health history and previous clinic and hospital visits across all practices
- If the patient has any procedure done at the hospital, this information should also be stored
  - Blood work, medical images (e.g. CT, Xray, MRI, colonoscopy scans, etc...), EKG scans, ECG scans, surgery information, and more
- Information on medical history, current vital and body statistics, medical procedures done, and SDOH can help identify abnormal trends in patient health for identification of patients at risk of readmission
- Information on each patient should be updated to a patient file after every visit. The overall database for the health care provider should be updated frequently (about 3-4 times per day) as they will collect large amounts of data frequently.

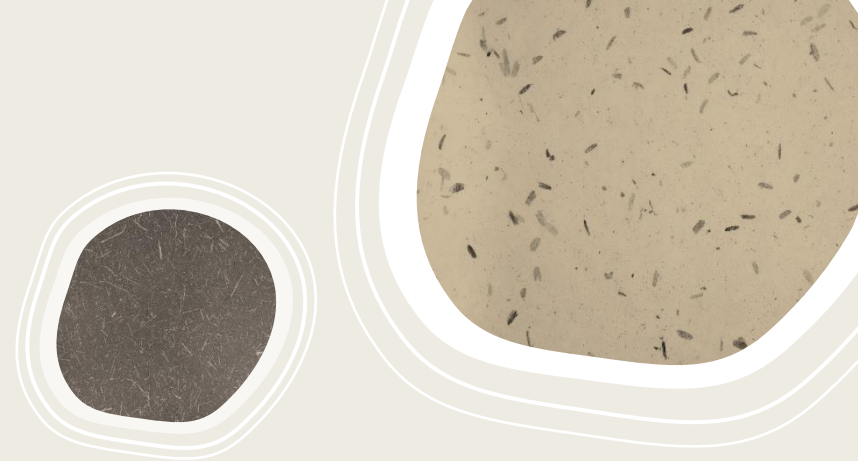
Data Analytics

- ML techniques can be used to identify if a patient is at risk of developing any diseases or medical conditions that would cause readmission
- ML can also be used to determine if there are certain patient populations that are readmitted more frequently than other populations
- Some techniques that could be used:
  - Computer vision on medical images
  - Regression, Neural Networks, Decision Trees/Random Forests, and Clustering for classification, pattern detection, etc...
- Similar analysis can be done to identify health and SDOH trends in patient populations that don't have high readmission rates
- Since it may take a while for patients to be readmitted to a hospital, the analytics should be performed and reviewed at least once every two weeks. However, analytics for in-patients and data related to vital signs should be reviewed daily. This will give enough time for data to be collected on patients that may have been readmitted without waiting too long to the point that the analysis is no longer helpful.

Big Data Solution

- Cloud Platform: Google Cloud Storage
  - Has multiple capabilities for storage and analysis
  - Has BigTable and BigQuery, allowing for both SQl and NoSQL usage
  - Can add firewall for added protection and security of patient data
- Storage Format and Compression Technique: Parquet
  - Useful for columnar data and nested data structures
  - Useful for compressing data in bulk
  - Ideally, data would be stored as a nested table with each entry in the larger data structure correlating to a single patient and each patient entry is formatted as a table (a nested table structure)
- Database and Query Execution Engine: BigTable and BigQuery
  - BigQuery can be used for SQL based analysis for the tabular data. Good for batch processing.
  - BigTable can be used for NoSQL based analysis which is useful for non-tabular data (e.g. medical images, sound data, unstructured data, etc…). It's also useful for graph-based analysis which can be used for clustering patient populations.
  - All the data will be initially added to a data lake before being structured and transferred to a data warehouse. BigTable will be used on the data lake while BigQuery will be used on the warehouse.
- Analytics/Data Science Platform:
  - VertexAI for notebook and coding (similar to a Jupyter notebook)
  - GCS also has DataFlow, DataProc, Pub/Sub, and can use Apache Spark for constructing ML pipelines

Data and Capacity Sizing



- Estimated data size per patient for one year: 100 MB
    - Includes all types of data format (e.g. text and images)
- Total data size initially: 1 PB
    - There's 500,000 patients, so within the first year there will be a total of 50 TB of data produced from all patients combined
    - To account for differing sizes of EHR data across patients due to differences in age (older patients will have more data than young), we estimate that for all past patient history plus history collected within the first year of implementing this big data solution, there will be 1 PB of data collected.
- Rate of data growth: 60 TB/year; 6% growth
    - Assuming all 500,000 patients grow at a rate of 100 MB per year
    - 10 TB to account for patients that may grow their data at faster rates (people with chronic health conditions) and to account for any new patients admitted
- HDFS sizing: 2.98125 PB
    - R = 3, C = ¾, S = 1 PB, T = 25%, G = 6%
- Capacity sizing:
    - No required data nodes since this is a cloud-based solution
    - 10 compute worker nodes should be enough for good efficiency and speed

Total Monthly Cost

COMPUTE	$7,645.26
Instances (Compute Engine)	$7,645.26
Service type Instances	
Instance-time 7300 Hours	N/A
Machine type custom, vCPUs: 30, RAM: 100 GB	$6,055.26
Instance-time 7300 Hours	N/A
GPU Model NVIDIA T4	N/A
Number of GPUs 1	$1,168.00
Local SSD 3x375 GB	$405.00
Boot disk type SSD persistent disk	N/A
Boot disk size (GiB) 10 GiB	$17.00
Number of Instances 10	N/A
Operating System / Software Free: Debian, CentOS, CoreOS, Ubuntu or BYOL (Bring Your Own License)	N/A
Provisioning Model Regular	N/A
Threads per core 2 threads per core	N/A
Enable Confidential VM service false	N/A
Add sustained use discounts false	N/A
Add GPUs true	N/A
Enable NVIDIA RTX Virtual Workstation false	N/A
Region Iowa (us-central1)	N/A
Committed use discount options 3 years	N/A

STORAGE	$43,770.16
Cloud Storage	$43,770.16
Service type Cloud Storage	
Data Transfer within Google Cloud 50 TB	$929.32
Replication type Default replication	$18,626.45
Total amount of storage 1 PB	$24,214.39
Location type Multi-region	N/A
Location United States (us)	N/A
Storage class Standard Storage	N/A
Source region North America	N/A
Destination region North America	N/A

Note that this estimate takes into account a discounted rate for 3 years of commitment. The price can also be lowered by reducing some of the storage and memory estimates.

Total Monthly Cost

DATA ANALYTICS	$593.43
BigQuery ML (BigQuery)	$593.43
Service type BigQuery ML	
Prediction 50 TB	N/A
Evaluation 5 TiB	N/A
Amount processed 50 TB	$593.43
Location type Region	N/A
Location Iowa (us-central1)	N/A
Model type DNN	N/A

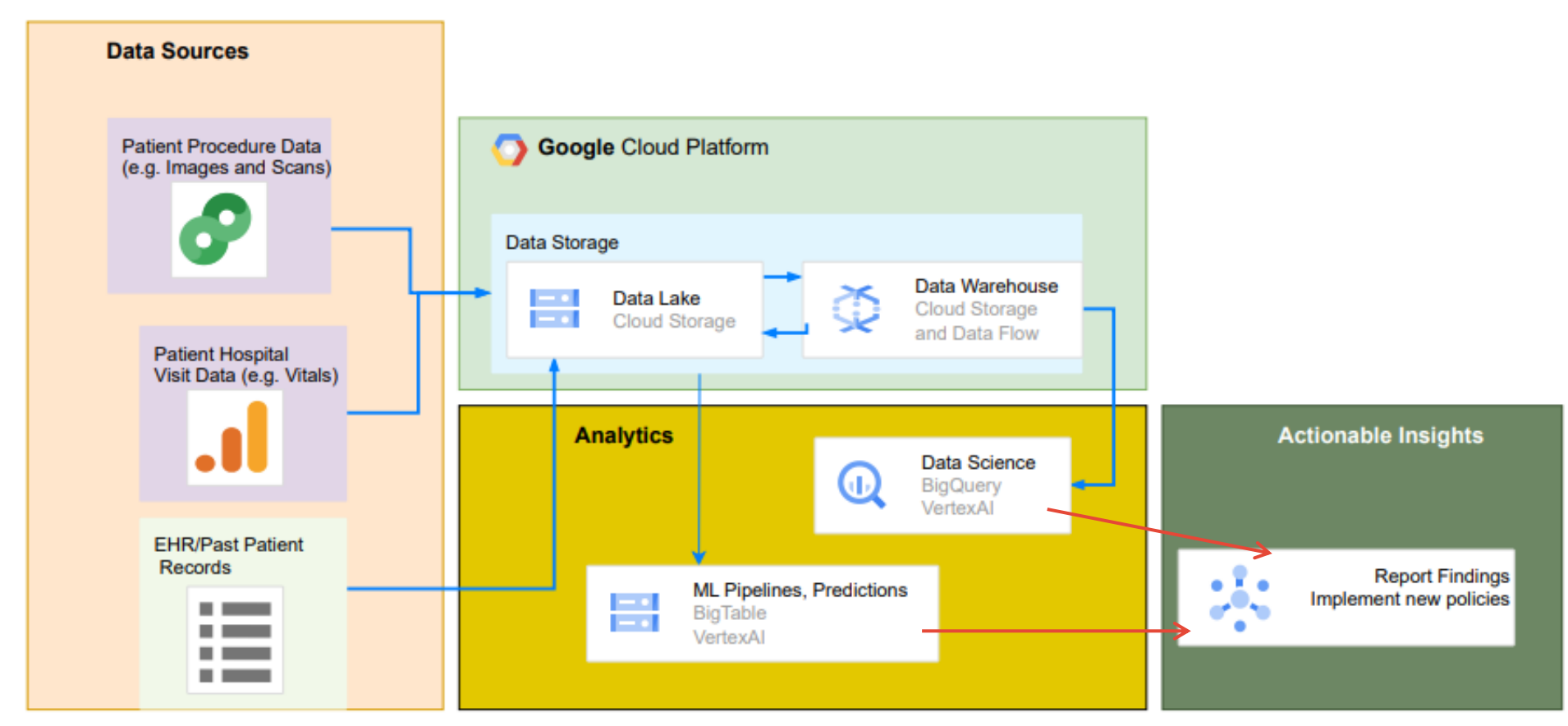
DATABASES	$14,861.06
Bigtable	$14,861.06
Service type Bigtable	
Node-time 7300 Hours	$2,847.00
Storage amount 50 TB	$4,749.75
Backup amount 500 TB	$7,264.32
Region Iowa (us-central1)	N/A
Number of Nodes 10	N/A
Storage type SSD	N/A
Committed use discount options 3 Years	N/A

AI & ML	$30.00
Vertex AI Pipelines (Vertex AI)	$30.00
Service type Vertex AI Pipelines	
Number of pipeline runs 1000	$30.00

Total Monthly Cost: $66,899.92
Estimating for BigQuery, BigTable, VertexAI, Storage, and Compute

In 2022, companies with 1,000 employees spent between $200,000 - $500,000 / month on storage

Reference Architecture Diagram



Sources (in addition to lecture notes)

Slides 3-4:

https://www.medicaladvantage.com/blog/ehr-vs-emr-what-is-the-difference/#:~:text=Scope%3A%20EMRs%20hold%20patient%20data,view%20a%20total%20patient%20record.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6669363/

https://www.cms.gov/about-cms/agency-information/omh/downloads/omh_readmissions_guide.pdf

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8101040/

Slides 5-6:

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8101040/

https://www.revelo.com/blog/google-cloud-vs-aws

https://www.upsolver.com/blog/the-file-format-fundamentals-of-big-data

https://bluexp.netapp.com/blog/gcp-cvo-blg-google-cloud-big-data-build-a-big-data-architecture-on-gcp

https://datacures.co/nosql-databases-healthcare-data-management/

https://cloud.google.com/data-science?hl=en

https://cloud.google.com/blog/topics/developers-practitioners/intro-data-science-google-cloud

https://www.nextech.com/blog/healthcare-data-growth-an-exponential-problem#:~:text=A%20first%20year%20patient%20may,addition%20to%20their%20text%20data.

Slides 7-8:

https://cloud.google.com/products/calculator?hl=en

https://www.cloudzero.com/blog/cost-of-cloud-computing/#:~:text=More%20specifically%2C%20in%20companies%20with,from%20%24600%2C000%20to%20%241.2%20million.

Slide 9:

https://medium.com/google-cloud/20-google-cloud-reference-architecture-to-start-your-gcp-architect-journey-2a85e8728507