2020

CSCI 49378: Lecture 10: Cloud Storage and Databases II

Bonan Liu
CUNY Hunter College

NYC Tech-in-Residence Corps

Cloud Storage and Databases II

Bonan Liu
Tech-In-Residence Member, Hunter College, CUNY
Spring 2020

Disclaimer

The content of this presentation is being provided for educational and informational purposes only. The views, thoughts, and opinions expressed in this presentation belong solely to the author, and not necessarily to the author's employer.

The content of this presentation is not endorsed by the author's employer.

Agenda

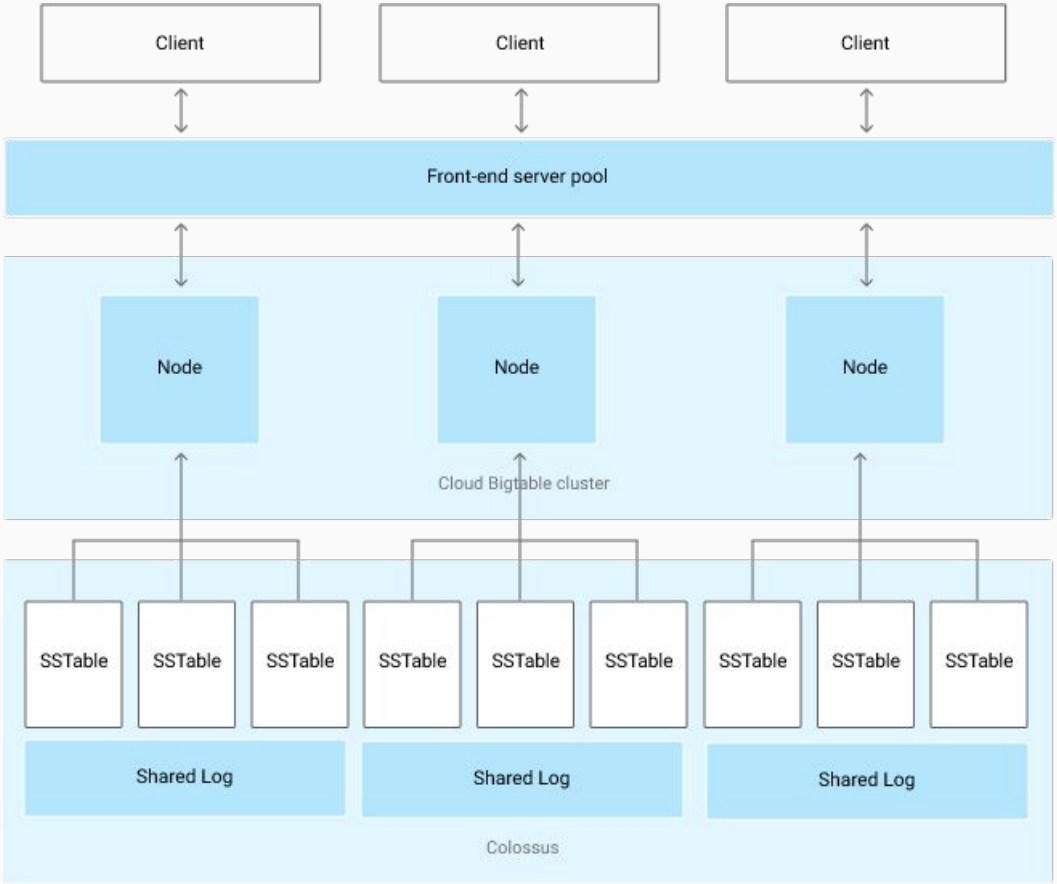
- Wide-Column Database: Cloud Bigtable
- Data Warehouse: Cloud BigQuery
- Review Assignment 3
- Review Assignment 4

Wide-Column Database: Cloud Bigtable

Cloud Bigtable is a wide-column database that could store O(thousands) columns and O(billions) rows. Cloud Bigtable could support petabytes of data.

- Scalable without interruption
- Replicated in multiple regions
- High performance:
  - Read: 1000 rows/second @ 6ms latency (SSD)
  - Write: 1000 rows/second @ 200ms latency (SSD)

Wide-Column Database: Cloud Bigtable



Wide-Column Database: Cloud Bigtable

Basic Elements:

- Table
- Row
- Column
- Column Family
- Cell
- Version

Wide-Column Database: Cloud Bigtable

Bigtable Schema:

- Only 1 index per table: row key
- Rows are stored by index alphabetically
- Columns are stored alphabetically within column family
- Empty cells do not take any storage space
- Each mutation take extra storage space
- Data compression: efficient when similar rows are stored together
- Data compaction

Wide-Column Database: Cloud Bigtable

Design the row keys:

- It’s important to properly design the row key because scanning row keys range are the only way to query
- Decide between string vs hashed string
- Decide between string vs timestamp
- Decide between random id vs sequential numeric ID
- Decide between new row vs new cell
- Decide between individual write vs batch write

Wide-Column Database: Cloud Bigtable

Cloud Bigtable Clusters:

- Instance
- Instance type
- Cluster
- Node
- Disk usage
- App profile

Data Warehouse: Cloud BigQuery

A data warehouse is a centralized data storage with easy data analysis tools.

Cloud BigQuery is a high-scalable serverless data warehouse which supports multiple data storage as data source. It also supports built-in BI and ML capabilities in the latest releases.

Data Warehouse: Cloud BigQuery

Three major operations with BigQuery:

- Importing data
  - Import/Export data
  - External data source
    - Bigtable
    - GCS
    - Drive

Data Warehouse: Cloud BigQuery

Three major operations with BigQuery:

- Querying data
  - On-demand Query
  - Batch Query
  - View
- Managing data
  - Manage datasets
  - Manage partitions

Demo: Running queries with BigQuery.

Readings

Bigtable: A Distributed Storage System for Structured Data.
https://static.googleusercontent.com/media/research.google.com/en//archive/bigtable-osdi06.pdf