CLOUD COMPUTING APPLICATIONS

Analytics in the Cloud

Prof. Reza Farivar

Data Analytics

- Business Intelligence (BI)
  - Set of technologies and processes that use data to understand and analyze business performance
- *Data Analytics:* the science of examining data to draw conclusions
  - A subset of BI
- Analytics is the discovery, interpretation, and communication of meaningful patterns in data
- extensive use of data, statistical and quantitative analysis, explanatory and predictive models, and fact-based management to drive decisions and actions

Advanced Analytics

- **Business Intelligence:**
  - “dark matter” of analytics—the necessary, but relatively simple, questions that will need to be answered frequently:
    - current inventory, number of customers, incoming/outgoing payments, average number of purchases per customer, etc.
  - Presented in dashboards, simple data plots and reports
- **Advanced Analytics:**
  - More complex statistical techniques and machine learning generate predictions and identify key performance indicators
  - Models of future consumer behavior based on past data
  - Fraud detection and recommender systems

OLTP vs OLAP

- OLTP
  - Online Transaction Processing System
  - typically involve most or all of the columns in a row for a small number of records
  - Using a database to run your business
  - RDBMS
    - Structured Data
    - SQL
    - Each query returns a small number of records
- OLAP
  - Online Analytical Processing System
  - read only a few columns for a very large number of rows
  - Using a database to understand your business
  - Data Warehouse
    - Structured Data
    - SQL
    - Each query covers many or all of the records
    - Typical query involves one column
- Different access patterns

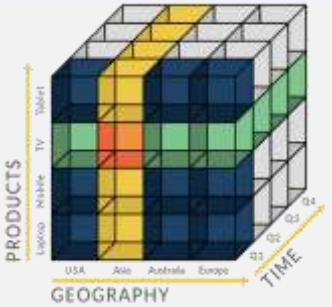
CLOUD COMPUTING APPLICATIONS

Analytics in the Cloud: Datacubes

Prof. Reza Farivar

Datacube Origins

- The OLAP cube grew out of a simple idea in programming: take multi-dimensional data and put it into what is known as a ‘2-dimensional array’ — that is, a list of lists
- A Datacube is a data structure
  - A sophisticated nested array
  - Compressions schemes
  - Data aggregation techniques when the cube outstrips the host's memory
- What if you have a massive dataset and want to run queries
  - Real technological constraints lead to the creation of the OLAP cube
  - Cache subsets of data within the nested array — and occasionally persist parts of the nested array to disk
- Today, ‘OLAP cubes’ refer specifically to contexts in which these data structures far outstrip the size of the hosting computer’s main memory — examples include multi-terabyte datasets and time-series of image data



Datacube Schemas

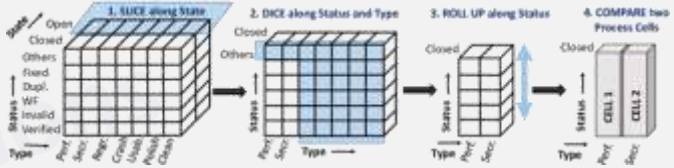
- OLAP cube requires that data teams manage complicated pipelines to transform data from an SQL database into cubes
- If you were working with a large amount of data, such transformation tasks could take a long time to complete, so a common practice would be to run all ETL (extract-transform-load) pipelines before the analysts came in to work.
- Using OLAP cubes in this manner also meant that SQL databases and data warehouses had to be organized in away that made for easier cube creation

Datacube Dimensional Modeling

- Early practitioners observed that certain access patterns occurred in every business
  - Kimball, Inmon and their peers
- They developed repeatable methods to turn business reporting requirements into data warehouse designs
  - designs that allow teams to extract the data they need in the formats they need for their OLAP cubes
- If you became a data analyst in the previous two decade, you had to “model” your data according to these best practices
  - Kimball dimensional modeling, Inmon-style entity-relationship modeling, or data vault modeling
  - Methods for organizing the data in the data warehouse to match the businesses' analytical requirements

Datacube Operations

- Slicing: the act of picking a rectangular subset of a cube by choosing a single value for one of its dimensions, creating a new cube with one fewer dimension
- Dicing: produces a subcube by allowing the analyst to pick specific values of multiple dimensions
- Drill Up / Down: allows the user to navigate among levels of data ranging from the most summarized (up) to the most detailed (down)
- Roll-up: A roll-up involves summarizing the data along a dimension
- Pivot: allows an analyst to rotate the cube in space to see its various faces



CLOUD COMPUTING APPLICATIONS

Analytics in the Cloud: Rise and Fall of Datacubes

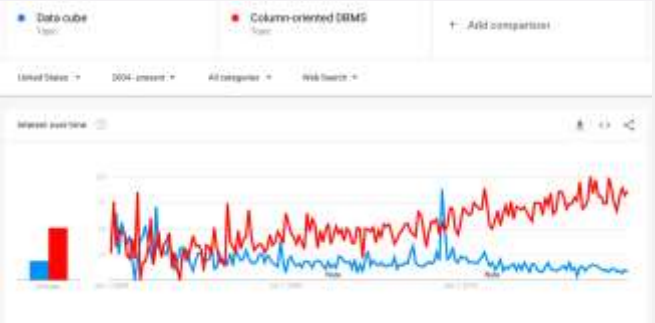
Prof. Reza Farivar

Datacube vs. Columnar RDBMS

- OLAP cubes traditionally known for extreme performance advantage over row-oriented RDBMS
  - Less important with recent advances in computers and columnar storage
- OLAP cubes demand that you load a subset of the dimensions you're interested in into the cube
- Columnar databases allow performing similar OLAP-type workloads at equally good performance levels without the requirement to extract and build new cubes
- Note: OLAP Datacubes typically offer richer analysis capabilities than RDBMSs, which are limited by the constraints of SQL
  - The main justification Datacubes are still relevant

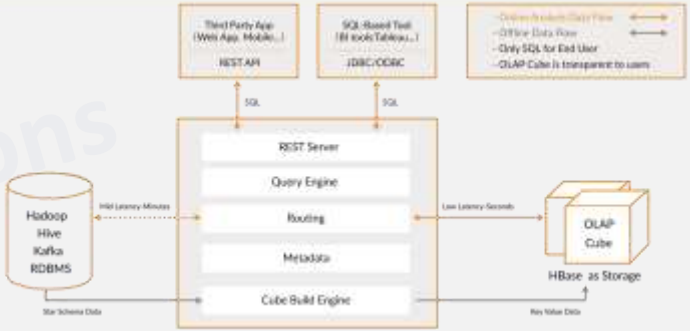
Current state

- Smaller companies are less likely to consider data-cube-oriented tools or workloads, and strict dimensional modeling has become less important over time
- Large tech giants (Google, Facebook, Amazon) have chosen columnar stores
  - Big Query, Redshift
- → One of the biggest shifts in data analytics over the past decade (2010 to 2020) is the move *away* from building Datacubes, to running OLAP workloads directly on columnar databases



Datacubes in the Future

- OLAP Datacubes typically offer richer analysis capabilities than RDBMSs, which are limited by the constraints of SQL
  - The main justification Datacubes are still relevant
  - OLAP cubes are being pushed upmarket
  - *We may return to them in the future*
- Example: Apache Kylin
  - Contributed by eBay in 2015
  - Build Datacubes on Hadoop and Spark
  - Utilizing HBase as Storage
  - Query billions of rows at sub-second latency
    - Identify a Star/Snowflake Schema on Hadoop
    - Build Cube from the identified tables
    - Query using ANSI-SQL and get results in sub-second, via ODBC, JDBC or RESTful API
- Druid, Apache Pinot (from Linkedin)
- Uber building a solution on Pinot + Presto



CLOUD COMPUTING APPLICATIONS

Analytics in the Cloud: Columnar Storage

Prof. Reza Farivar

History

- **Column-stores.** In recent years, there has been renewed interest in so-called *column-oriented systems*, sometimes also called *column-stores, a.k.a. Columnar Storage*
- MonetDB
- VectorWise → Ingres VectorWise → Actian Verctor
- C-Store → Vertica
- SybaseIQ

Columnar Storage

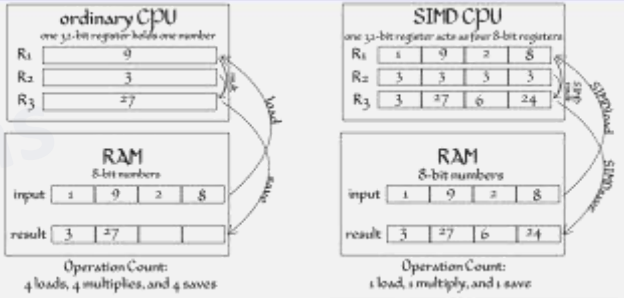
- Traditionally, databases stored records in rows, similar to how a spreadsheet appears
  - e.g. this could include all information about a customer or a retail transaction.
- Retrieving data the traditional way required the system to read the entire row to get one element
- With columnar storage, each data element of a record is stored in a column
- With this approach, a user can query just one data element, such as gym members who have paid their dues, without having to read everything else in that entire record, which may include each member's ID number, name, age, address, city, state, payment information, and so on.
  - Reading the same number of column field values for the same number of rows requires a fraction of the I/O operations and uses a fraction of the memory that would be required for processing row-wise blocks
  - For example, suppose a table contains 100 columns. A query that uses five columns will only need to read about five percent of the data contained in the table.

Columnar Storage

- Column-oriented stores are a good fit for analytical workloads that compute aggregates, such as finding trends, computing average values, etc.
  - Read Optimized
- Processing complex aggregates for when records have multiple fields, but some of them have different importance and are often consumed together
- Note: Column-oriented databases should not be mixed up with wide column stores, such as BigTable or Hbase
  - Data represented as a multidimensional map
  - Columns are grouped into column families (usually storing data of the same type)
  - Inside each column family, data is stored row-wise
  - This layout is best for storing data retrieved by a key or a sequence of keys.

Hardware Optimization: Cache and SIMD

- Disk access pattern
  - One SSD page is 4KB~8KB
  - Row-store: When reading a page, a small number of similar column fields from different rows are loaded
  - Column Store: All the read page are relevant column fields
- Reading multiple values for the same column in one run significantly improves cache utilization and computational efficiency
- On modern CPUs, vectorized instructions (SIMD) can be used to process multiple data points with a single CPU instruction



Hardware Optimization: Compression

- Storing values that have the same data type together (e.g., numbers with other numbers, strings with other strings) offers a better compression ratio
  - Lower information entropy resulting in higher compression
  - We can use different compression algorithms depending on the data type and pick the most effective compression method for each case
- Compression can be automatic by the engine
- Columns shrink and grow independently

```
CREATE TABLE loft_deep_dive (
    aid    INT         ENCODE LZO
   ,loc    CHAR(3)     ENCODE BYTEDICT
   ,dt     DATE        ENCODE RUNLENGTH
);
```

Updates in Columnar Stores

- Update performance in a columnar database is poor
  - Go to every column in order to update one ‘row’
- Many modern columnar databases limit the ability to update data after it is stored
  - Example: Google Dremel paper explains the system as append-only structure
  - No longer a limit of Google BigQuery
  - Performance might be slow
- Redshift
  - Each block is 1MB, blocks are immutable
  - Clone blocks on write to not cause fragmentation
  - Samll writes (~1-10 rows) has similar cost to larger writes (~100K rows)

Metadata

- To reconstruct data tuples, which might be useful for joins, filtering, and multirow aggregates, we need to preserve some metadata on the column level to identify which data points from other columns it is associated with
- Example: AWS Redshift storage nodes have 2.5~3X more storage attached than advertised
  - Used internally for metadata

Column Store File Format

- During the last several years, likely due to a rising demand to run complex analytical queries over growing datasets, we've seen new column-oriented file formats
- Apache Parquet, Apache ORC, RCFile, as well as column-oriented stores, such as Apache Kudu, ClickHouse
  - Parquet: an open source file format for Hadoop
    - Hive, Pig, Impala, Spark
    - Parquet stores nested data structures in a flat columnar format
  - ORC, the Optimized Row Columnar format

CLOUD COMPUTING APPLICATIONS

Analytics in the Cloud: Modern Data Warehouses

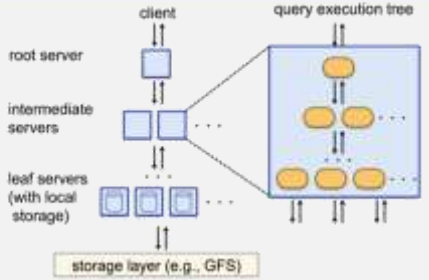
Prof. Reza Farivar

Modern Data Warehouse Architecture

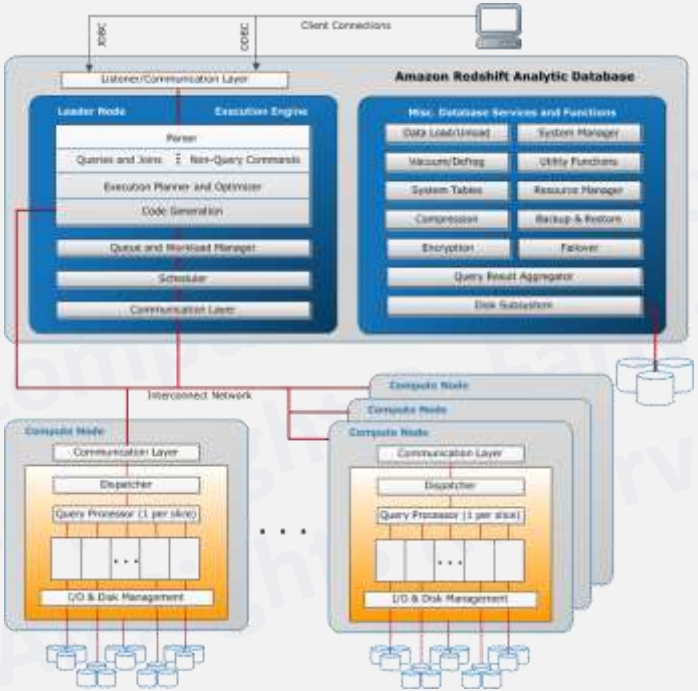
- **Cloud**
  - access to near- infinite, low-cost storage
  - improved scalability
  - Outsourcing of data warehousing management and security to the cloud vendor
  - Pay per use
- **Massively parallel processing (MPP)**
  - Dividing computing operations to execute simultaneously across many separate computer processors
  - Like Sharding
- **Columnar storage**
- **Vectorized processing**

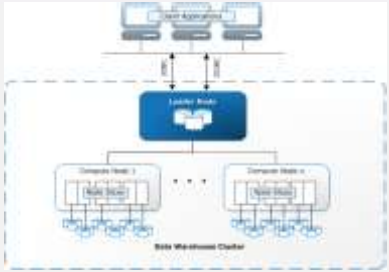
Columnar-based Data Warehouses

- Column Store, MPP, Cloud based
- MariaDB with InfiniDB
  - For reference: row based regular engine for OLTP: InnoDB
- PostgreSQL
  - Citus cstore_fdw
- Google BigQuery
  - Based on Google Dremel, paper published in 2010
- AWS Redshift
  - Based on an older version of PostgreSQL
    - PostgreSQL 8.0.2
    - Originally developed by ParAccel
    - Some PostgreSQL features that are suited to smaller-scale OLTP processing, such as secondary indexes and efficient single-row data manipulation operations, have been omitted to improve performance



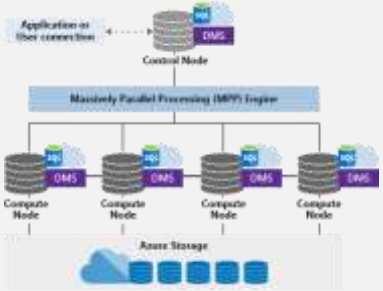
Redshift Architecture





Microsoft Azure Synapse

- Azure Synapse Analytics
  - Formerly Azure SQL Data Warehouse
  - SQL Analytics: Complete T-SQL based analytics
    - SQL pool (pay per DWU provisioned)
    - SQL on-demand (pay per TB processed)
  - Spark: Deeply integrated Apache Spark
  - Integration with Power BI



CLOUD COMPUTING APPLICATIONS

Analytics in the Cloud: Data Lake

Prof. Reza Farivar

Data Lake

- Data Warehouses cannot accommodate unstructured big data projects
  - Petabytes of data in structured, semi-structured and unstructured forms
  - Semi-structured and unstructured data: JSON, XML, Log files, Natural Language, Images, video, etc.
    - Social media sites, mobile phones, Internet of Things (IoT) devices, and many other sources, including shared data sets
  - Structured data typically collected from enterprise applications
- Data Lake: a new type of data repository for storing massive amounts of raw data in its native form, in a single location
  - "A large body of water, into which new water streams from many channels, and from which samples are taken and analyzed"
  - Solution to a growing problem: the need for a scalable, low-cost data repository that allowed organizations to easily store ***all*** data types and analyze that data to make evidence-based business decisions
- The initial data lakes were deployed on premises, mostly using open source tools from the Apache Hadoop ecosystem
- Modern data lakes combine the power of analytics with the flexibility of big data models and the agility and limitless resources of the cloud

Components of a Modern Data Lake

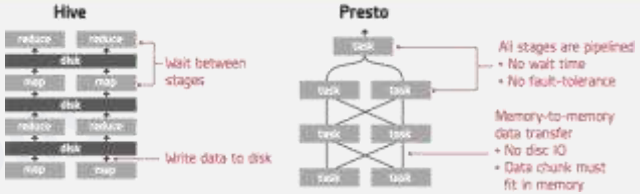
- Object Storage to store all data types
  - Big Data: Azure Data Lake Storage
- Move Data
  - AWS Data Pipeline
  - Azure Data Factory
- Data Lake Schema Discovery:
  - A fully managed service that serves as a system of registration and system of discovery for enterprise data sources
  - AWS Glue
  - Azure Data Catalog
- SQL Exploration and Query
  - Apache Presto
  - AWS Athena
- Lake Formation
  - AWS Lake Formation
  - Azure Data Share

Discovery

- Data Lake Discovery Services
- Data Crawler
- Metadata extraction → Schema → Catalog
- ETL workloads
- Apache Atlas
- AWS Glue
  - Serverless
  - Data Sources: Amazon Redshift, Amazon S3, Amazon RDS, and Amazon DynamoDB
  - AWS Glue Data Catalog: Crawls your data sources, identifies data formats, and suggests schemas and transformations
  - AWS Glue ETL:
    - AWS Glue provides a number of built-in preload transformations that let ETL jobs modify data to match the target schema
    - Automatically generates code to execute data transformations and loading processes for more complex, custom ETL transformations
  - ETL jobs on a fully managed, scale-out Apache Spark environment
- AWS Data Pipeline: Focus on data transfer
- Azure Data Catalog: Discovery
- Azure Data Factory: ETL
- Google Cloud Data Catalog, Dataflow

Data Lake Exploration and Query

- Directly run SQL queries on the data lake
- No need to setup intermediary databases or data warehouses
- Apache Presto
  - In-memory distributed SQL query engine
  - Optimized for star schema joins
    - 1 large Fact table and many smaller dimension tables
  - Interactive Hive on Steroids
- Aws Athena
  - Managed Serverless Presto
- Azure Data Lake Analytics
- Apache Spark SQL
  - IBM Cloud SQL Query




```
@a  =
    SELECT * FROM
        (VALUES
            ("Contoso", 1500.0f),
            ("Woodgrove", 2700.0f)
        ) AS
            D( customer, amount );
OUTPUT @a
    TO "/data.csv"
    USING Outputters.Csv();
```

Example Azure Data Lake Analytics U-SQL query on a Data Lake output goes to a csv file on the data lake

Cloud-based Data Lake Automation

- Automated tools to orchestrate the data transfer, discovery, ETL and analytics steps
- AWS Lake Formation
  - Glue
  - Athena
  - Redshift Spectrum
  - EMR
    - Apache Zeppelin or EMR Notebooks
- Azure Data Share

CLOUD COMPUTING APPLICATIONS

Analytics in the Cloud: Other Analytics Services

Prof. Reza Farivar

Serverless Analytics

- Azure Analysis Services
  - Built on SQL Server Analysis Services
  - Tabular models only
  - Partitions, perspectives, row-level security, bi-directional relationships, and translations
- AWS Redshift Spectrum
- AWS Athena

Search-based Analytics

- Full Text Search
- Search analytics
- ELK Stack
  - ElasticSearch
  - Logstash
  - Kibana
- Search
  - AWS CloudSearch
    - Based on Apache Solr search engine
      - Solr uses MapReduce
      - Hadoop was born from Solr
  - Azure Cognitive Search

Big Data Analytics

- Open Source Big Data Analytics
  - Managed Hadoop, Spark, Hive
    - AWS EMR
    - Azure HDInsight: Hadoop, Spark, Kafka, HBase, Storm, etc.
    - Azure Databricks
- Streaming Analytics
  - AWS
    - Kinesis Data Analytics
    - Amazon Managed Streaming for Apache Kafka (Amazon MSK)
  - Azure
    - Stream Analytics

Azure HDInsight Ecosystem

Graphical BI Tools

- Visualization
  - Tableau
  - AWS QuickSight
  - Azure PowerBI