Data Engineering in the Cloud

Hybrid Transactional Analytical Processing

Xuemao Zhang
East Stroudsburg University

January 18, 2025

Outline

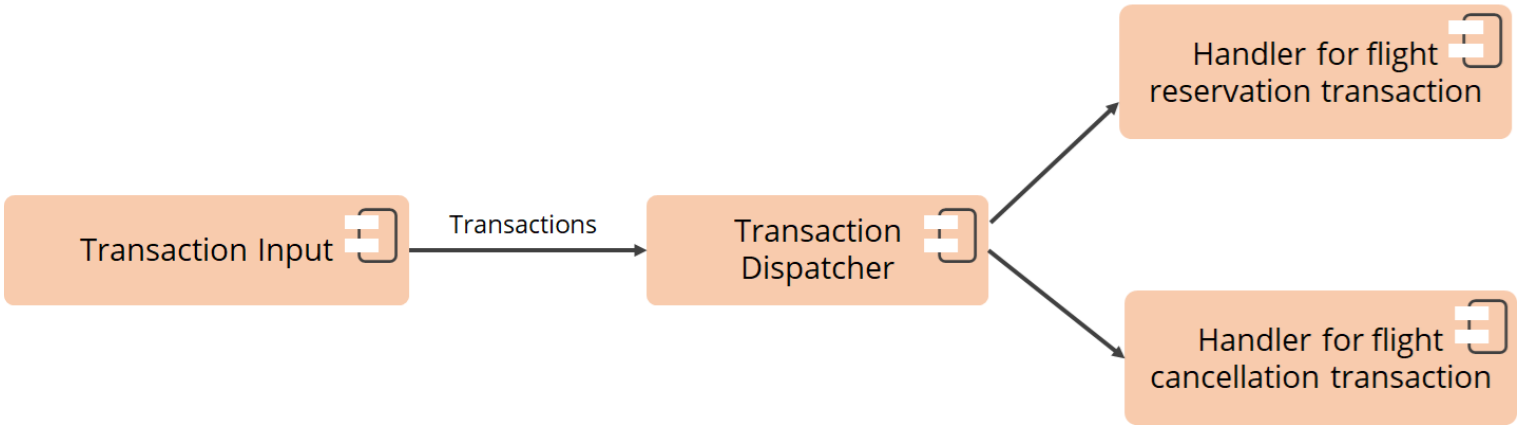
- Transactional Processing
- Analytical Processing
- Hybrid Transactional and Analytical Processing
- NoSQL Databases
- Azure Synapse Link and Cosmos DB

Transactional Processing

- While a transaction is in progress, transactional processing focuses on looking at filtered data and performing operations such as update, remove, and insert.
- Transactional Processing:
  - ACID Properties: Ensures Atomicity, Consistency, Isolation, and Durability of transactions.
- OLTP (Online Transaction Processing): Manages transaction-oriented applications.
  - Focus: High transaction throughput, maintaining data integrity in multi-access environments.
- OLAP (Online Analytical Processing): Supports complex queries and data analysis.
  - Focus: Aggregation, data mining, and multi-dimensional analysis.

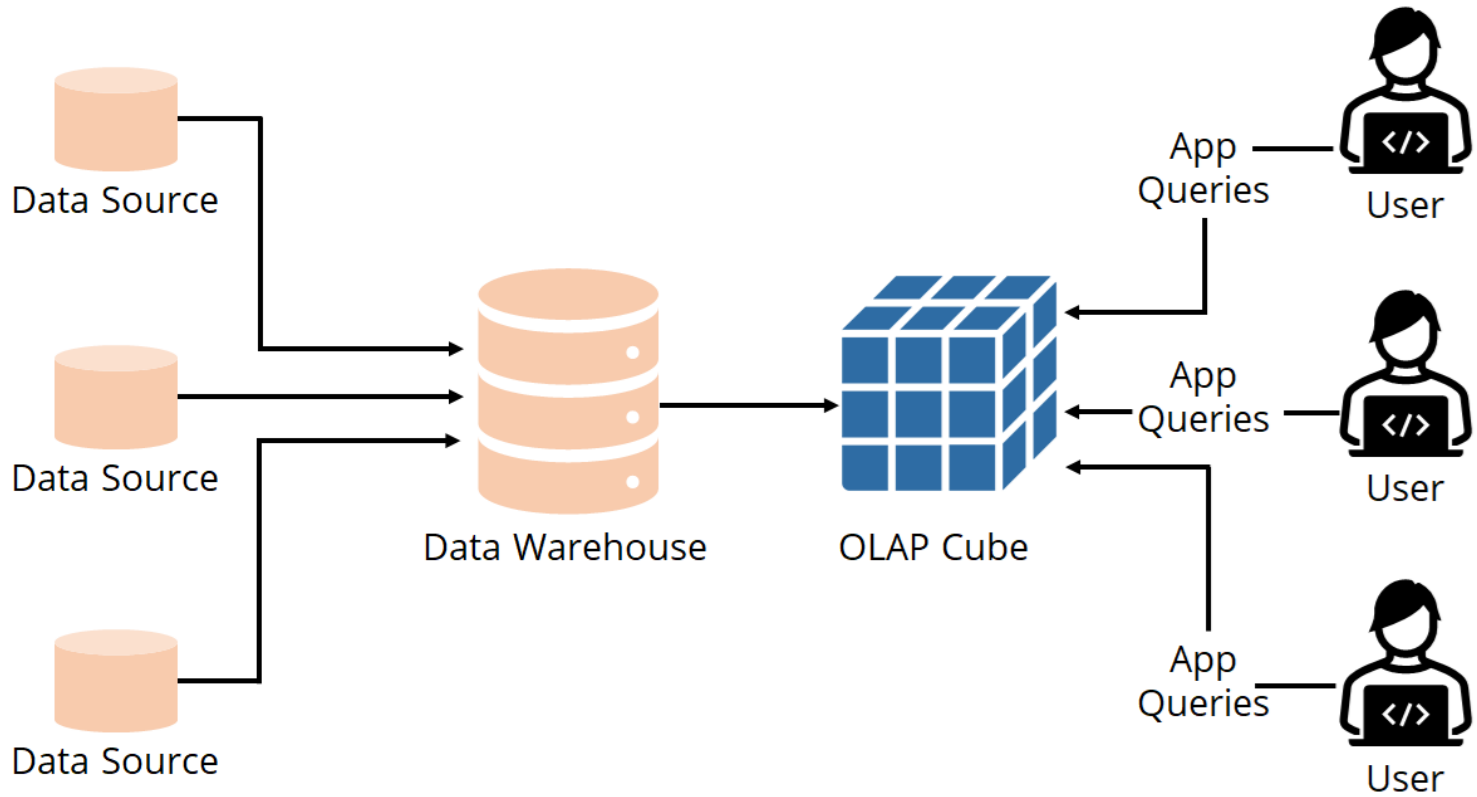
Transactional Processing

- Transactional processing queries undergo standardized queries, and the data updates are fast.
- Example of Transactional Processing System



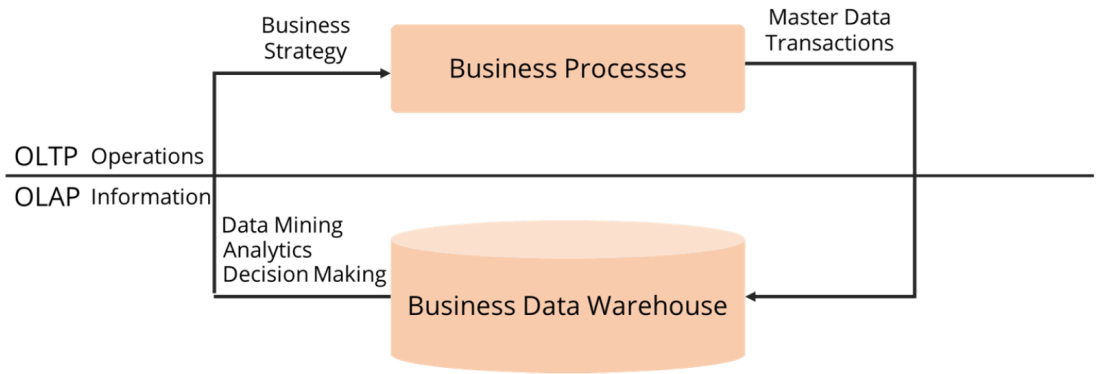
Analytical Processing

- Analytical processing works with a large amount of historical data gathered from various sources and applies complex queries to it.



Hybrid Transactional and Analytical Processing

- Several business applications separate transactional and analytical processing, and the data is stored on different infrastructures.
- These infrastructures are known as OLAP and OLTP for systems that work with historical data and systems that work with operating data, respectively.



Hybrid Transactional and Analytical Processing

- OLTP + OLAP = HTAP (Hybrid Transaction or Analytical Processing)
- HTAP systems can handle real-time analytics and transactional workloads simultaneously, allowing for faster insights and decision-making.
  - Examples: Modern financial systems, IoT applications, real-time fraud detection.
- By leveraging these technologies, organizations can ensure efficient transactional processing while gaining real-time analytical insights, driving informed decision-making and enhancing operational efficiency.

Hybrid Transactional and Analytical Processing

Example Scenario: Consider an e-commerce platform that handles thousands of transactions per minute.

- OLTP Role: Manages transactions like placing orders, updating inventory, and processing payments.
- OLAP Role: Analyzes historical sales data, customer behavior, and inventory levels for strategic decisions.
- HTAP Role: Provides real-time insights into sales trends and inventory status, enabling dynamic pricing and instant restocking decisions.

Implementation in Azure

Implementation in Azure:

- Azure Cosmos DB: Supports transactional processing with ACID transactions for single partition key operations.
- Azure Synapse Analytics: Provides OLAP capabilities for data warehousing and complex data analysis.
- HTAP Solutions: Utilizing Cosmos DB with Synapse Link for near real-time analytics on operational data.

NoSQL Databases

- NoSQL is a form of unstructured storage.
- NoSQL stands for "Not Only SQL" and refers to a class of database management systems that do not adhere strictly to the traditional relational database model.
- NoSQL databases provide a mechanism for storage and retrieval of data that is modeled in means other than the tabular relations used in relational databases.

NoSQL Databases

Types of NoSQL databases

- 1. Key-Value: Key-Value stores are the simplest type of NoSQL databases. They store data as a collection of key-value pairs where a key is used to uniquely identify the value. The value can be any type of data, such as strings, JSON, BLOBs, etc.
  - Simple, fast, and scalable for basic key-value pair data.

- 2. Document-Based Stores: organize data in documents, which are usually formatted in JSON, BSON, XML, or similar. Each document is a self-contained unit of data with a unique key, and can contain nested structures.
  - Flexible and versatile for semi-structured data in document formats.

NoSQL Databases

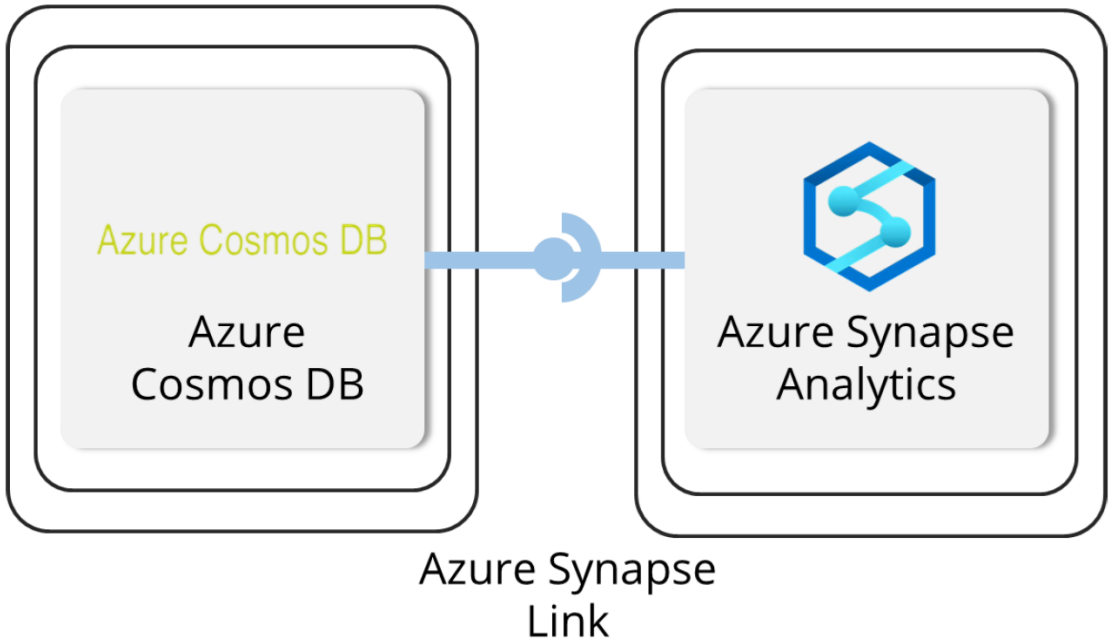
Types of NoSQL databases

- 3. Column-Based Stores (Wide-Column Stores): store data in columns rather than rows. They group related data into column families, which allows for efficient read and write operations, especially in cases of large datasets.
  - Efficient for large-scale data analytics, time-series data and Real-time analytics

- 4. Graph-based stores use graph structures with nodes, edges, and properties to represent and store data. They are particularly effective for applications where relationships between data points are key.
  - Ideal for complex relational data, such as social networks and Fraud detection

Azure Synapse Link and Cosmos DB

- Azure Synapse Link is a Microsoft Azure integrated analytics solution that provides customers with choices for incorporating Spark and SQL into data warehousing and other analytical needs.
- Power BI, Cosmos DB, and Azure ML are a few Azure services that can link with Synapse.
- Azure Synapse provides security over the data lake, unified management, and monitoring. We can use Azure Synapse Studio for building complete analytical solutions.



Cosmos DB

- Microsoft Azure's Cosmos DB is a powerful, globally distributed, multi-model NoSQL database service.
- Multi-Model: Cosmos DB supports various data models, enabling you to use the most appropriate model for your application. It includes support for:
  - Graph Data: Using the Gremlin API.
  - Document Data: Using the MongoDB API and SQL API.
  - Key-Value Data: Using the Table API.
- Global Distribution: Cosmos DB is designed to be globally distributed, allowing you to replicate your data across any Azure region.
  - Setting up a globally distributed database can be complex and time-consuming. It involves careful planning for data replication, consistency, failover, and disaster recovery.
  - Cosmos DB simplifies this process by providing a fully managed service that takes care of these challenges, allowing you to focus on building your application rather than managing the infrastructure.

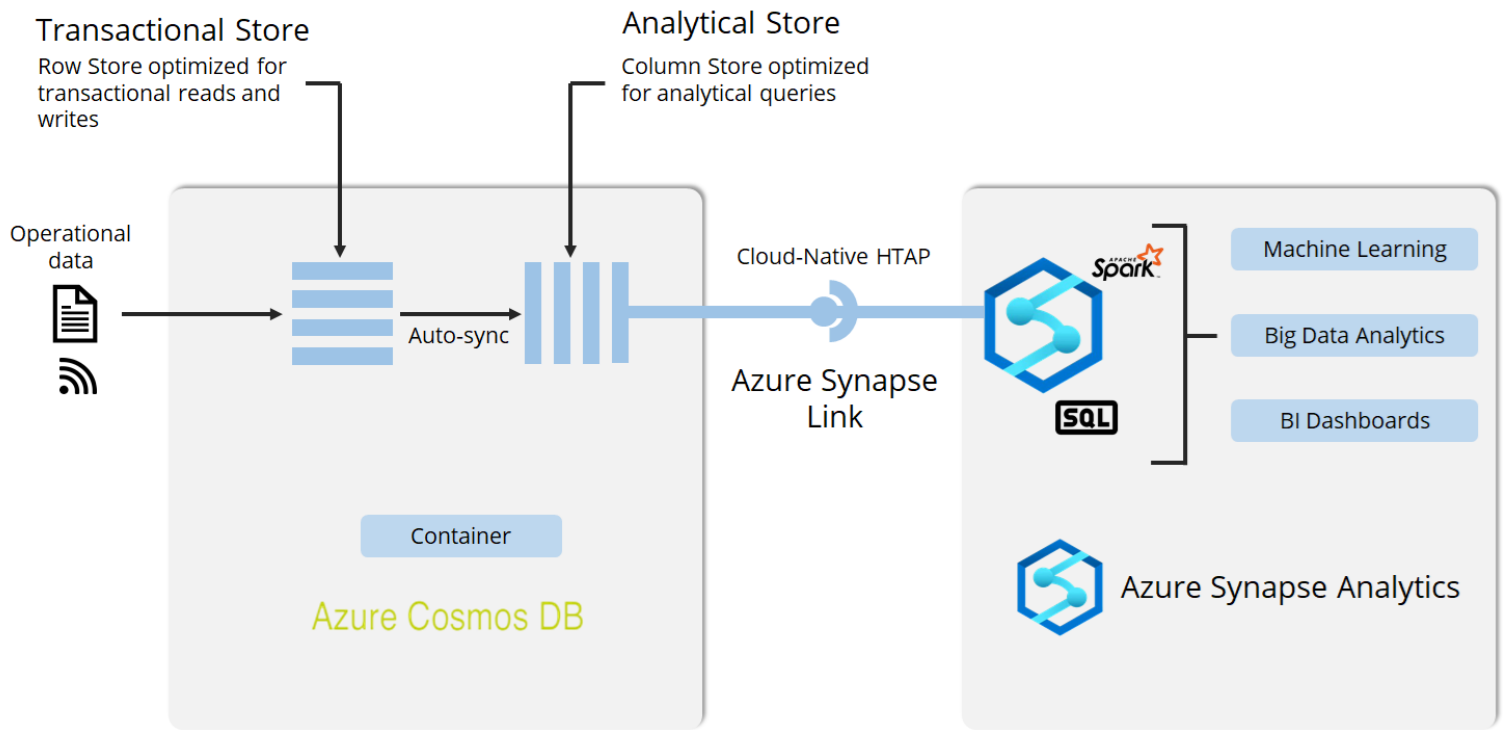
Cosmos DB

Advantages:

- Elastic Scalability: Cosmos DB allows you to independently and elastically scale throughput and storage. You can start small and grow as your application's needs increase.
- Global Distribution: Easily replicate your data across multiple regions to ensure high availability and low-latency access for users worldwide.
- High Availability: With a 99.99% availability SLA, Cosmos DB ensures your applications remain up and running.
- Platform as a Service (PaaS): As a PaaS offering, Cosmos DB abstracts much of the complexity associated with setting up and managing a globally distributed database.

Implement Configuration of Azure Synapse Link with Cosmos DB

- Microsoft Azure's Synapse Link for Cosmos DB is a cloud native hybrid transactional and analytical processing service.
- Synapse Links enable organizations to consume and generate insights from their data in real time using the query analytics tools of their choice.



Implement Configuration of Azure Synapse Link with Cosmos DB

- 1. Transactional Store:
  - Row Store: Optimized for transactional reads and writes, ensuring that operational data is quickly and efficiently processed.
  - Operational Data: Data generated from various operational sources, such as applications, IoT devices, and user interactions, is stored here.

- 2. Analytical Store:
  - Column Store: Optimized for analytical queries, providing efficient storage and retrieval for large-scale data analytics.
  - Auto-Sync: Automatically synchronizes data from the transactional store to the analytical store, ensuring that the data is up-to-date without manual intervention.

Implement Configuration of Azure Synapse Link with Cosmos DB

- 3 Azure Synapse Link:

Cloud-Native HTAP: Facilitates the seamless integration between Azure Cosmos DB and Azure Synapse Analytics, enabling real-time analytics on transactional data.

- 4 Azure Synapse Analytics:

Query Analytics Tools: Allows users to utilize various analytics tools, such as Apache Spark and SQL, for data processing and analysis.

Steps to Configure the Synapse Link

- Step 1: Enable the Synapse Link for Azure Cosmos DB account.
- Step 2: Create a Cosmos DB container with analytical stores enabled
- Step 3: Change TTL for the analytical store (optional)
- Step 4: Connect Synapse workspace with Cosmos DB
- Step 5: Use Synapse Spark to perform queries on the analytical store
- Step 6: Use Serverless SQL pool to perform queries on the analytical store
- Step 7: Use serverless SQL to analyze and visualize data in Power BI
- Check the lab in the next lecture.

License