snowflake

CLOUD ANALYTICS CONFERENCE

LONDON

@SnowflakeDB #CloudAnalytics17

Bringing Your Data Together in the Cloud

Todd Beauchene

Global Alliances Architect, Snowflake Computing

“Data! Data! Data!
I can't make bricks without clay.”

-Sherlock Holmes

Agenda

- Cloud Data Ecosystem
- Data Sources
- Methodologies
- Data Integration Solutions
- Conclusion

Cloud Data Ecosystem

Data Sources

- Enterprise apps
- Corporate
- Web
- Mobile
- IoT

Data Integration

talend

kafka

Data Warehouse

Business Intelligence & Analytics

Spark

looker

tableau

Data Sources

Data Sources

On-Premises

- Typically backed by a local transactional database
- All data lives within the firewall
- Customer has full access to all data and system

Cloud

- Typically backed by a cloud database (i.e. RDS)
- Can run in customer VPC
- Typically offers fewer options than on-premises

SaaS

- Typically data is only available via API
- Outside of customer firewall or VPC
- Customer has very little control over handling of data

Real World Example: Consolidated Dashboard

Challenges	Solutions
• Long-term project with high-level goals	• Agile project with focused, short-term goals
• Diverse data sources	• Dedicated schema in EDW
• Different refresh cycles	• Daily ETL Process
• Inconsistent results	• Data quality checks within ETL

Methodologies

Methodologies

Bulk Loading – Trunc and Load

- Runs at regular intervals
- Full dataset loaded during each run and existing data is purged
- Least efficient option, but very simple to manage
- High data volumes every run
- More commonly used for dimension tables

Daily Differentials

- Runs during nightly ETL window
- Requires change data capture to identify changed rows
- Generally consists of a series of steps where each depends on the previous steps
- Must include logic to handle slowly changing dimensions

Methodologies

Insert-only – Date-based

- Extracts data by date range to eliminate need for CDC
- Simplified processing
- Commonly used for fact tables
- Changes to data from previous periods require deletion of all data for the given range

Database Replication

- Generally runs in near-real-time
- Requires a tool that is tightly integrated with the source database
- Schemas must match between source and destination

Methodologies

Batch Processing

- Generally used when data is being pushed from the source
- Batch frequency depends on the volume and velocity of the data
- Requires automated process to load batches into the data warehouse.

Streaming

- Generally used for high volume data
- Event-based rather than row-based
- Often requires micro-batching of data for load into relational database
- Raw data must usually be transformed to support analytics

Data Integration Solutions

Data Integration Solutions

Custom Code

- Flexible but complex
- Leverages in-database processing
- Challenging to manage and maintain

ETL

- Simplified data transformation with no code
- Built-in dependency and error handling
- Reduces data volumes within EDW

ELT

- Leverages benefits of ETL while shifting data processing to EDW
- Requires tight integration between Data Integration and EDW
- Raw and transformed data in one place

Data Integration Solutions

On-Premises

- Customer owns hardware and software install/configuration
- Don’t have to deal with firewall to access local sources

Cloud

- Customer owns software install/configuration but not hardware
- Can run in customer VPC to provide direct access to data within VPC or behind firewall

SaaS

- Fully managed by service provider
- Configurable options vary by solution
- Must find secure ways to access data not stored inside firewall

Conclusion

Cloud Data Warehousing Best Practices

- Leverage the scalable compute layer to do the bulk of the data processing
- Isolate load and transform jobs from queries to prevent resource contention
- Eliminate physical datamarts by leveraging a scalable data platform
- QA is key, make sure all changes made to data integration tasks are tested before they roll to production
- When migrating it is important to convert one source at a time