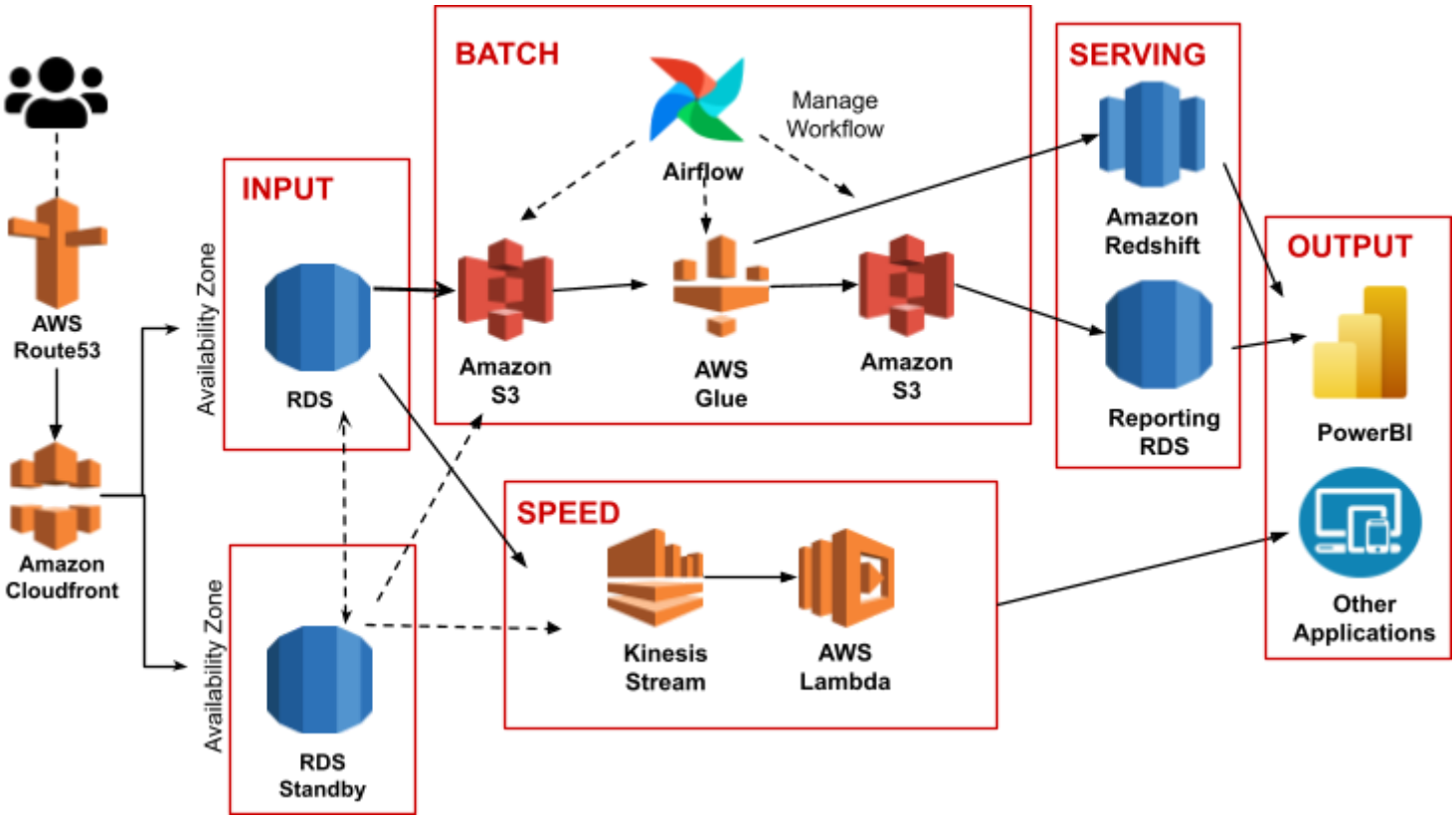
AWS Architecture



Components	Role	Details
Amazon RDS (Relational Database Service)	Hosts the primary relational database.	Used to host a primary OLTP database. It provides managed database services, automatic backups, and high availability.
Amazon S3 (Simple Storage Service)	Storage for data files and backups.	Store batch data files in S3 for processing. Also, used S3 for storing database backups.
AWS Glue	ETL (Extract, Transform, Load) service.	AWS Glue can be used to transform and prepare data before loading it into the RDS database.
Amazon Kinesis Data Streams	Ingests and manages real-time streaming data.	Kinesis Data Streams allows you to collect and process real-time data streams.
AWS Lambda	Serverless computing for executing code in response to events.	Used to stream data for real-time processing.

Amazon Redshift	Data warehousing solution for analytics and reporting.	It provides a fast and scalable way to analyze large datasets and run complex queries.
Reporting RDS	Hosting a secondary relational database for reporting purposes.	Used a separate RDS instance for reporting to avoid impacting the primary OLTP database's performance.

Data Loading Strategies:

Batch Processing:

1. **Data Ingestion to S3:** Batch data files are stored in Amazon S3. Periodic batch processes or ETL jobs read data from these files.
2. **AWS Glue ETL:** AWS Glue performs ETL operations on batch data. Transformation and preparation of data occur before loading it into the secondary RDS database.
3. **Backup Data to S3:** Transformed files are stored in Amazon S3 for backup.
4. **Loading into RDS:** Processed data is loaded into the secondary RDS database. This can be done in scheduled batches, ensuring that large volumes of data are efficiently processed.

*Airflow is used to managed the workflow

Real-Time Processing:

1. **Amazon Kinesis Data Streams:** Real-time streaming data is ingested and managed through Kinesis Data Streams. Events or data are sent to the stream in near real-time.
2. **AWS Lambda for Real-Time Processing:** AWS Lambda functions are triggered in response to events from Kinesis Data Streams. These functions process and analyze real-time data.

Characteristics of the System:

Resilience:

- **Database Backups:** Automatic backups provided by Amazon RDS ensure data recoverability.
- **Separate Reporting RDS:** A separate reporting RDS instance enhances resilience by isolating reporting activities from the primary OLTP database.
- **Enable Cross-Region Replication:** Create read replicas in different AWS regions to enhance fault tolerance. This ensures that data is still available even if an entire region becomes unavailable.
- **Set up S3 Cross-Region Replication:** Replicate data across different AWS regions for additional redundancy and disaster recovery.

High-Performance:

- **Amazon Redshift:** Amazon Redshift is used for analytics and reporting, providing high-performance query capabilities.
- **AWS Lambda for Real-Time Processing:** AWS Lambda enables real-time processing without the need for maintaining server infrastructure, contributing to high performance.

Security:

- **Amazon RDS Security:** Amazon RDS provides security features such as encryption, access control, and network isolation.
- **S3 Security:** S3 security features ensure the confidentiality and integrity of batch data.
- **IAM Roles for Lambda:** AWS Identity and Access Management (IAM) roles are used to control access to Lambda functions.

Scalability:

- **Kinesis Data Streams:** Scales elastically to handle varying volumes of real-time data.
- **Serverless Architecture with AWS Lambda:** Scales automatically based on the incoming event load.