Delta Live Tables (DLT) Concepts and Examples

Introduction:

Delta Live Tables (DLT) is a framework in Databricks designed to simplify the development, deployment, and monitoring of reliable data pipelines. It provides declarative pipeline definitions, data quality rules, and automatically manages dependencies, retries, and lineage tracking.

Core Concepts:

1. DLT allows defining ETL pipelines as declarative tables using Python or SQL.
2. Supports Live and Streaming tables for both batch and streaming ingestion.
3. Includes built-in data quality enforcement using 'expectations'.
4. Handles automatic data lineage and table dependency tracking.
5. Optimized with Delta Lake for ACID transactions and versioning.

Example SQL Definition:

```
CREATE LIVE TABLE customers_cleaned AS SELECT id, name, email FROM
STREAM(live.customers_raw) WHERE email IS NOT NULL;
```

Example Python Definition:

```
@dlt.table(comment="Cleans raw customer data") def customers_cleaned(): return (
spark.readStream.table("customers_raw") .filter("email IS NOT NULL") )
```

Materialized Views and Stream Tables:

DLT allows creating materialized views using SQL syntax and continuous streaming tables that maintain up-to-date data using event time and watermarks.

Monitoring and Quality Enforcement:

DLT pipelines include automatic monitoring of data quality and job status through the Databricks UI. You can define expectations for validating data using built-in syntax.