SNOWFLAKE DATA ENGINEERING DECISION TREE - 12/02/2021 - v1.02

Nick Akincilar - 12/2/2021

START

Is the data from 3rd party? — Yes → Does the 3rd party use Snowflake?

No ↓

Is it already in Snowflake?

Does the 3rd party use Snowflake? — Yes → **Use Live DataShare** (Direct or MarketPlace)

Is it already in Snowflake? — No → **INGEST DATA** / Yes → **TRANSFORM DATA**

INGEST DATA
- Real-Time Streaming
- Batch

TRANSFORM DATA

Less than 400K files Per Day — No → Consistant Workload Size?

Less than 400K files Per Day:
- Yes → **Snowpipe** (Serverless & Event based)
- Option → **Kafka Connector** (Serverless & Event based)

Consistant Workload Size?
- Yes → **Standard task** (Time Based, User Managed Compute) → **Virtual Warehouse**
- No → **Serverless Task** (Time Based, Auto Scaling Compute)

Can Schema Change or need raw data for future reference?
- Yes → Source as **Semi-Structured** data(Json, Parquet, Avro & etc.) & ingest into RAW table as a VARIANT
- No → Source as **Flat**(CSV, TSV) Files & ingest directly into Staging Table with matching columns

Optional ↓ Use **Schema detection**

** In Private Priview **

Need to Keep Track of Changes in Data CDC?
- Yes → Use **STREAMS** tables for Transformations
- No → Use **Standard** Tables for Transformations

Are Complex functions not possible via SQL required?

Performing ML Scoring, External API Lookups, Sentiment Analysis & etc..

- Yes → Requires external API lookups
  - Yes → Use **External Functions** (Fast)
  - Option →
  - No → Use **Java/Python UDFs** (Faster)
- No → Use **ANSI-SQL** (Fastest)

Add **Clustering Keys** or **Materialized Views** for Better Performance

** Optional ***

Querying specific values like IP Address, order number & etc with small resultsets?

Is it Big Data? Table > 1TB
- Option
- Yes → Is the data spot lookup low latency use case?
  - Yes → **** Optional **** Enable **Search Optimization** if Amount of changes/Inserts do not modify bulk of the entire data.
  - No ↓
- No → **Create TARGET Tables**

Create TARGET Tables

Table has PII data?
- I don't know → Use **AUTOMATIC DATA CLASSIFICATION** to apply PII tags **** In Private Priview ****
- No → **Create DATA WAREHOUSE Tables/Views**
- Yes → Use **TAGGING** to Identify Tables & Columns with PII data → Create **DATA MASKING** Policies using TAGS → Create/Apply **DYNAMIC DATA MASKING** → Does PII data used for Analytics?
  - No → **Create DATA WAREHOUSE Tables/Views**
  - Yes → Use **DATA ANONIMIZATION** (*Masks PII data but keeps the analyical value*) **** In Private Priview ****