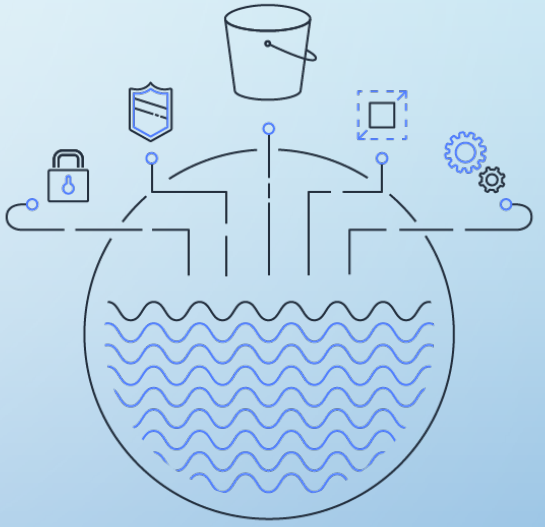
LARGE-SCALE DATA PREPARATION FOR MACHINE LEARNING MODELS

FLOCON 2023

MATTHEW SPITZER, PHD; ROSALIE BAKKEN, PHD; AND MS. JENNIFER MARR

CHARACTERISTICS OF INCOMING DATA

- Amount of data
  - Significant volumes - trillions of records across multiple sources
  - Petabyte-scale storage and searchability required
  - 24x7x365 accumulation of new data
- Types of data
  - Data sources are highly structured but unrelated
    - Individual source systems send data independently of other sources
    - Data is primarily network flow records and endpoint metadata
    - System-generated vs manually generated data
  - Arrival velocity
    - Recently accumulated data is ingested as frequently as hourly
  - Data velocity
    - Time-based table partitions are used to optimize search parameters and costs

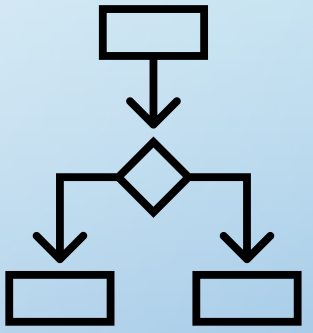


CHALLENGES OF SOURCE DATA

- Staleness due to operational process failures or lack of awareness
  - Source system owners must maintain currency of data
  - Requires effort to drive long-term resolutions with owners
- Data collection processes
  - Collaborate with system owners to create stable data generation processes
- Differences between display data and raw data
  - System owners are used to working with displayed data, not raw data
  - Foreign key values may be present in raw data, which will require ingestion of the reference tables to correctly query/interpret the raw data
- Vendor implementation affects data generation
  - Data may be formatted according to RFC standards, or it may be a proprietary format exclusive to a vendor
  - Consider when to use raw data vs. proprietary data when appropriate

IMPLICATIONS FOR PROCESSING

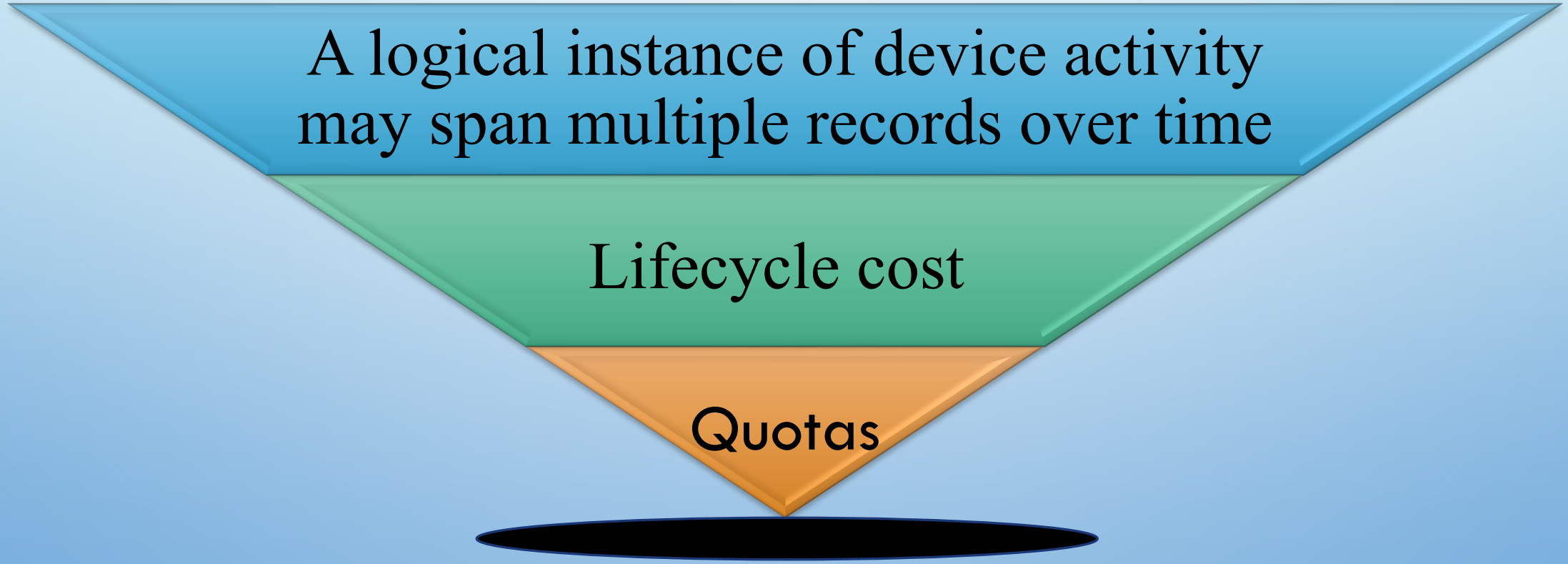
- Look at incoming data holistically
  - Macro and micro view of the lake (tributaries to streams to rivers to a lake)
  - Understand the data and source
  - Are there lookups that need to be included?
- Constraints
  - Resource: Bandwidth, number of files, storage (on premise)
  - Quotas: Cloud based resource limits vary by provider
- Maximizing efficiency per dollar spent
  - False sense of urgency to fill the lake
  - Intimate data knowledge leads to more efficient development
- Deliberate pause to perform a true evaluation
  - Breathe



EVALUATING THE DATA

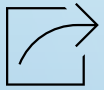
- Evaluate samples of raw data
  - Sample must be representative of population for data source
  - Identify available fields and values in each source
  - Evaluate field accuracy against high-volume noise
    - Identify hygiene issues for remediation
    - Identify transformation rules
    - Identify guidelines for interpretation of field values
- Determine expected volume requirements
  - Constraint impacts (on-premise and cloud)

CAUTIONARY TALES

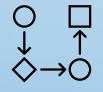


PROCESSING METHODOLOGY

- Source to Target mapping
  - Retain context of field values even after processing



- Implement reliable transfer mechanism for data from the source systems
  - Include monitoring to alert on gaps in expected data velocity

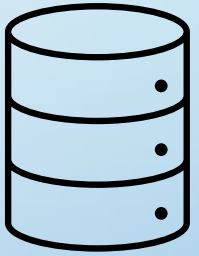


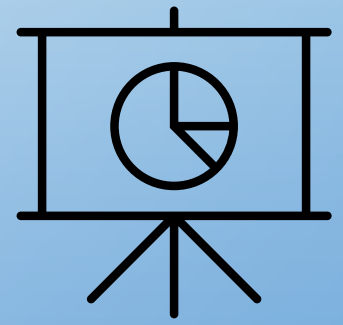
- Process incoming data – parse, filter, clean, transform
  - Implement error notifications to actively alert on data quality issues

- Commit processed data to the lake

DATA PREPARATION CHALLENGES

- Accuracy and Completeness
  - Does the processed data match the raw values in context?
  - Do derived values reflect a true representation of the raw data?
  - Are there missing fields and/or values that must be reconciled?
- Disparate source fusion
  - Can we reliably "stitch" tables together based on related fields?
  - How do differences between sources affect the ability to reliably obtain results?
- Scalable statistical analysis
  - How do the results of the analysis correlate to the data?
  - Do the results make sense?
  - How does feature extraction improve results?





STATS ANALYSIS & FEATURE EXTRACTION

- Feature extraction (FE) - collate specific columns of interest
  - Initial observations of the raw data showed connection fragments are represented in multiple records
    - Aggregate fragments into cohesive "conversations"
  - Without FE, we would have scanned 3.2PB vs. 350TB
    - $ savings + query time savings + effort savings

- Statistical analysis (increase confidence in data integrity)
  - Frequency analysis – shows distribution of data points
  - Patterns – identify patterns in connection (port, bytes, direction of connection – internal/external)
  - Voluminous data – FE enabled smaller query size and more efficient processing
  - May reveal avenues for further investigation
- Reproducibility and deviation identification

ACCOMPLISHMENTS AFTER IMPLEMENTING FEATURE EXTRACTION (USE CASE SPECIFIC)

350 TB (out of 3.3 PB) scanned / 88% reduction in volume of data scanned
$1,750 cost per query (assuming no cache, vs $16k on the raw data)
Reduced query execution time for all queries – improved efficiency
Streamlined, clean data relationships from feature extraction effort
Easily pivot viewpoints within the data (flexible perspectives when observing results)
Positioned for ongoing reduction of scanned data (daily updates)
High degree of confidence in data integrity as we move to data modeling efforts
Data is well positioned for use case analysis

USE CASE: FEATURE EXTRACTION ENABLING LRC IDENTIFICATION/EVALUATION

- Long Running Connections
  - i.e. those lasting longer than 20 days, in this example
- Data volumes sent over this connection may be variable
- Connections may come in different forms which may indicate C2C
  - For example: SSH or RDP
- Difficulties exist in tracking long running connections
  - What is the baseline for traffic to endpoints?
  - How to distinguish legitimate from malicious activity?
    - Statistical analysis in previous steps may help correlate possible malicious activity
- Implication and learnings for future investigations and modeling

QUESTIONS