4.2 — Project Briefing

Data Processing at Scale

2026-03-27

Data Processing at Scale — Day 4

Course website

https://vbergeron.github.io/data-processing-at-scale/

This presentation

https://vbergeron.github.io/data-processing-at-scale/4.2-project-briefing.pdf

What will you build?

Pick a dataset. Build a system.
Defend your architectural choices.

Guidelines

- Let's pretend it is big data — keep the data volume manageable
- Tech stack is your choice
- Explain, don't describe

Evaluation

- 100% project presentation (separate day)
- A working system processing the chosen dataset
- Understanding of architectural choices and trade-offs
- Ability to answer questions about internals

GH Archive

- Every public GitHub event since 2011 (push, star, issue, PR…)
- 5–15 GB compressed (1 week to 1 month subset)
- Direct HTTP download, no authentication

#1 — Star-Farming Ring Detection

- **Pitch:** Detect coordinated starring rings, score inflated repos
- **Difficulty:** 

#2 — Bus Factor Analysis

- **Pitch:** Contributor concentration in PySpark vs DuckDB — find the crossover
- **Difficulty:** 

NYC Taxi

- Every taxi and for-hire trip in NYC (pickup, dropoff, fare, tip)
- 10–20 GB Parquet (6–12 months yellow taxi)
- Direct HTTP download, no authentication

#3 — Urban Pulse: Event Detection

- **Pitch:** Model the city's rhythm from taxi flow, detect real-world events
- **Difficulty:** 

#4 — Fare Anomaly Detection

- **Pitch:** Per-route statistical baselines, temporal drift with adaptive windows
- **Difficulty:** 

Wikimedia EventStreams

- Real-time stream of all edits across Wikipedia and sister projects
- 2–5 GB for a 24–48h recording
- Open SSE endpoint, no auth — also available via Kafka

#5 — Vandalism Detection

- **Pitch:** Score edits for vandalism likelihood using session-based behavioral signals
- **Difficulty:** 

#6 — Knowledge Graph Evolution

- **Pitch:** Incremental maintenance of link graph metrics as edits arrive
- **Difficulty:** 

Stack Exchange

- Complete dump of all Stack Exchange sites since 2008
- 20 GB uncompressed (Stack Overflow subset)
- Free download from Internet Archive

#7 — Incremental Analytics

- **Pitch:** Maintain answer quality scores without recomputation, reason about monotonicity
- **Difficulty:** 

#8 — Distributed Faceted Search with Embedded Lucene

- **Pitch:** Lucene indexes inside Spark `mapPartitions`, BM25 + facets
- **Requires:** Spark (Scala) + Lucene — JVM embedding
- **Difficulty:** 

GDELT

- Global news monitoring: who did what to whom, 100+ languages
- 15–45 GB CSV (1–3 months of events)
- Direct HTTP download, updated every 15 minutes

#9 — Probabilistic Event Tracking

- **Pitch:** HyperLogLog, Count-Min Sketch, t-digest — compare exact vs approximate
- **Difficulty:** 

#10 — Escalation Early Warning

- **Pitch:** Detect deteriorating country-pair relations from rolling Goldstein scores
- **Difficulty:** 

US Stocks / SEC EDGAR

- Daily OHLCV (open, high, low, close, volume) for all US stocks
- Quarterly financial statements
- 10–15 GB combined
- Kaggle (account) or SEC EDGAR (no auth)

#11 — Time-Series Analytics

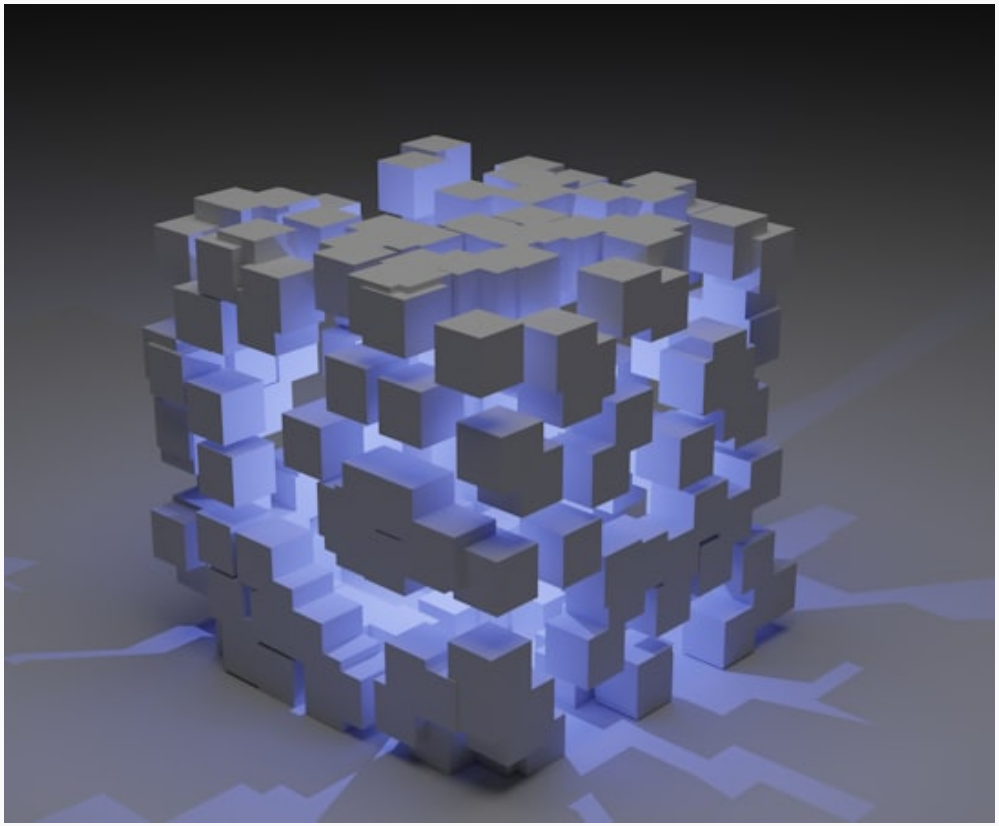
- **Pitch:** Rolling indicators + materialized view cascade, measure write amplification
- **Difficulty:** 

#12 — Financial Statement Pipeline

- **Pitch:** Temporal as-of join between daily prices and quarterly filings
- **Difficulty:** 

Ethereum + Sourcify

- Every transaction and token transfer on Ethereum + verified contract ABIs
- 10–20 GB (3–6 months of transactions)
- Free via `ethereum-etl` CLI or BigQuery



#13 — Decoded On-Chain Analytics

- **Pitch:** Join transactions with ABI data, decode function calls, analyze gas
- **Difficulty:** 

#14 — DeFi Streaming Monitor

- **Pitch:** Replay token transfers in Flink, detect volume spikes, handle chain reorgs
- **Difficulty:** 

OpenSky Network

- ADS-B flight tracking worldwide — position every second
- 10–20 GB (1–3 months of state vectors)
- Free registration required, bulk download

#15 — Flight Density Analytics with Embedded H3

- **Pitch:** Hexagonal spatial indexing with H3, flight density heatmaps
- **Difficulty:** 

#16 — Flight Diversion & Anomaly Detection

- **Pitch:** Detect circling, U-turns, rapid descent from live ADS-B streams
- **Difficulty:** 

Reddit (Pushshift)

- Full archive of all comments and submissions since 2005
- 15–40 GB compressed (1–3 months of comments)
- Free via Academic Torrents

#17 — Toxicity Scoring with Embedded ONNX

- **Pitch:** Distributed ML inference with an embedded ONNX model
- **Difficulty:** 

#18 — Narrative Cross-Pollination Tracker

- **Pitch:** Detect topic diffusion across subreddits using Bloom filters
- **Difficulty:** 

Spotify Million Playlist

- 1M user-created playlists, 66M track entries, 2M unique tracks
- 5 GB JSON (full dataset is manageable)
- Free from AICrowd (account required)

#19 — Playlist Similarity Engine

- **Pitch:** Exact Jaccard vs MinHash/LSH approximate similarity at scale
- **Difficulty:** 

#20 — Genre Boundary Detection

- **Pitch:** Artist co-occurrence graph, community detection, genre-bridging
- **Difficulty:** 

OpenStreetMap

- The full map of Earth, collaboratively edited — every changeset recorded
- 5–10 GB (changesets + 1–3 months of replication diffs)
- Free download, no authentication

#21 — Collaborative Edit CRDT Analysis

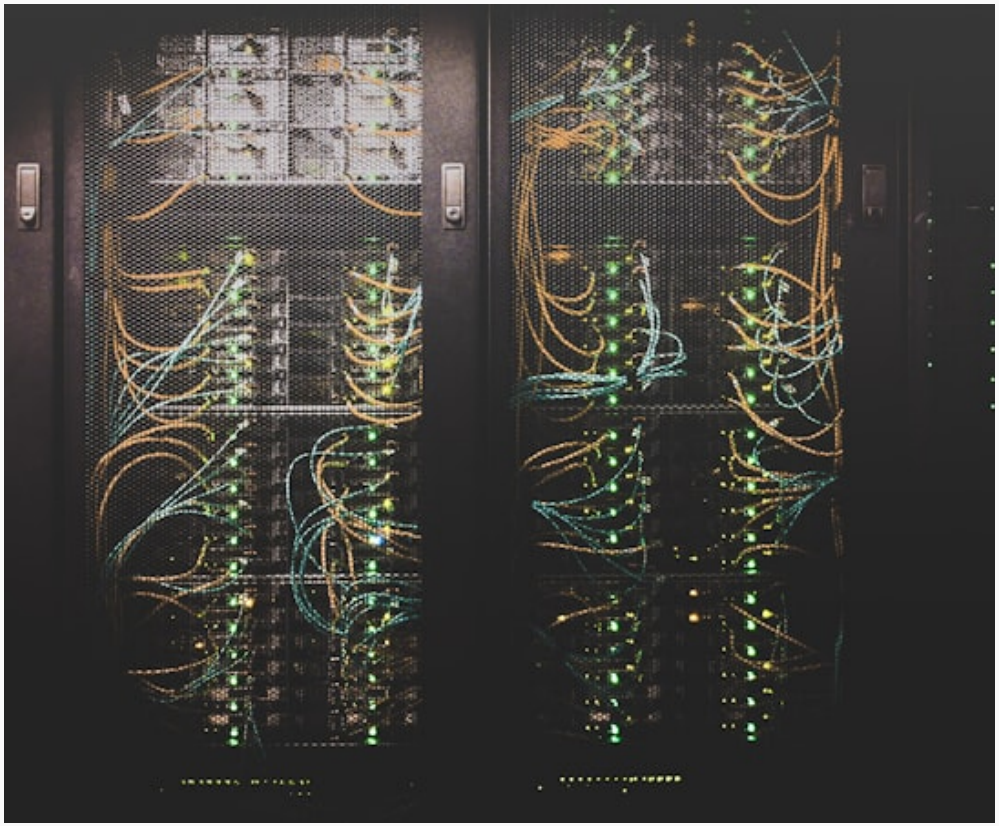
- **Pitch:** Model editing as a lattice, detect conflicts, reason about convergence
- **Difficulty:** 

#22 — Consistency Validator with Embedded Prolog

- **Pitch:** Embedded Prolog engine for rule-based consistency checks on map tiles
- **Difficulty:** 

Common Crawl

- Monthly crawl of the entire public web (2–3 billion pages per crawl)
- 10–20 GB Parquet (columnar index for one crawl)
- Free from S3 (requester-pays) or direct HTTP



#23 — Web-Scale Search with Embedded Lucene

- **Pitch:** Sharded Lucene indexes inside Spark, cross-shard query merging
- **Requires:** Spark (Scala) + Lucene — JVM embedding
- **Difficulty:** 

#24 — Link Graph Incremental Processing

- **Pitch:** Maintain PageRank across crawl versions using diffs
- **Difficulty:** 

Binance (Live Websocket)

- Real-time trades and order book updates for all trading pairs
- 5–15 GB (24–72h recording for replay)
- Public websocket, no API key required

#25 — Real-Time Portfolio Tracker

- **Pitch:** Transactionally consistent portfolio state, exactly-once 2PC sink
- **Difficulty:** 

Binance (Live Websocket)

#26 — Order Book Reconstruction & Arbitrage

- **Pitch:** Live order book in Flink state, triangular arbitrage detection
- **Difficulty:** [redacted]

Citi Bike

- Every Citi Bike trip in NYC since 2013 + real-time station status
- 10 GB CSV (trip history) + live JSON feed
- Free download, no authentication

#27 — Demand Hybrid Pipeline

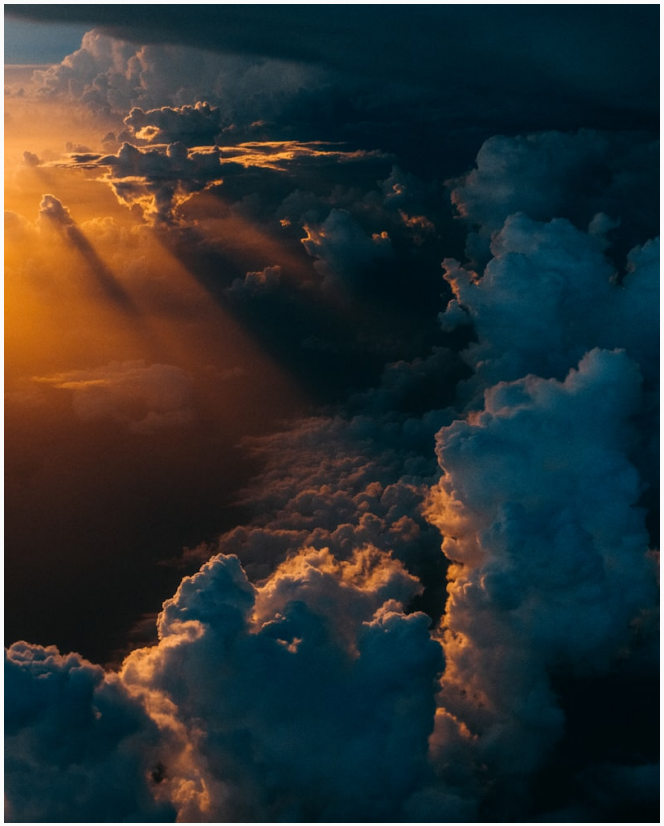
- **Pitch:** Batch history in Spark + live station feed in Kafka/ClickHouse
- **Difficulty:** 

#28 — Fleet Rebalancing Stream

- **Pitch:** Maintain station bike counts, detect imbalance, reason about non-monotonic state
- **Difficulty:** 

NOAA Weather

- Daily weather from 10,000 stations worldwide since 1929
- 3–10 GB (10–30 years of data)
- Free download, no authentication



#29 — Extreme Weather Event Attribution

- **Pitch:** Per-station baselines, compound event detection, trend analysis
- **Difficulty:** 

#30 — Global Weather Correlation

- **Pitch:** Pairwise cross-station correlations, O(n²) optimization strategies
- **Difficulty:**