snowflake

YOUR DATA, NO LIMITS

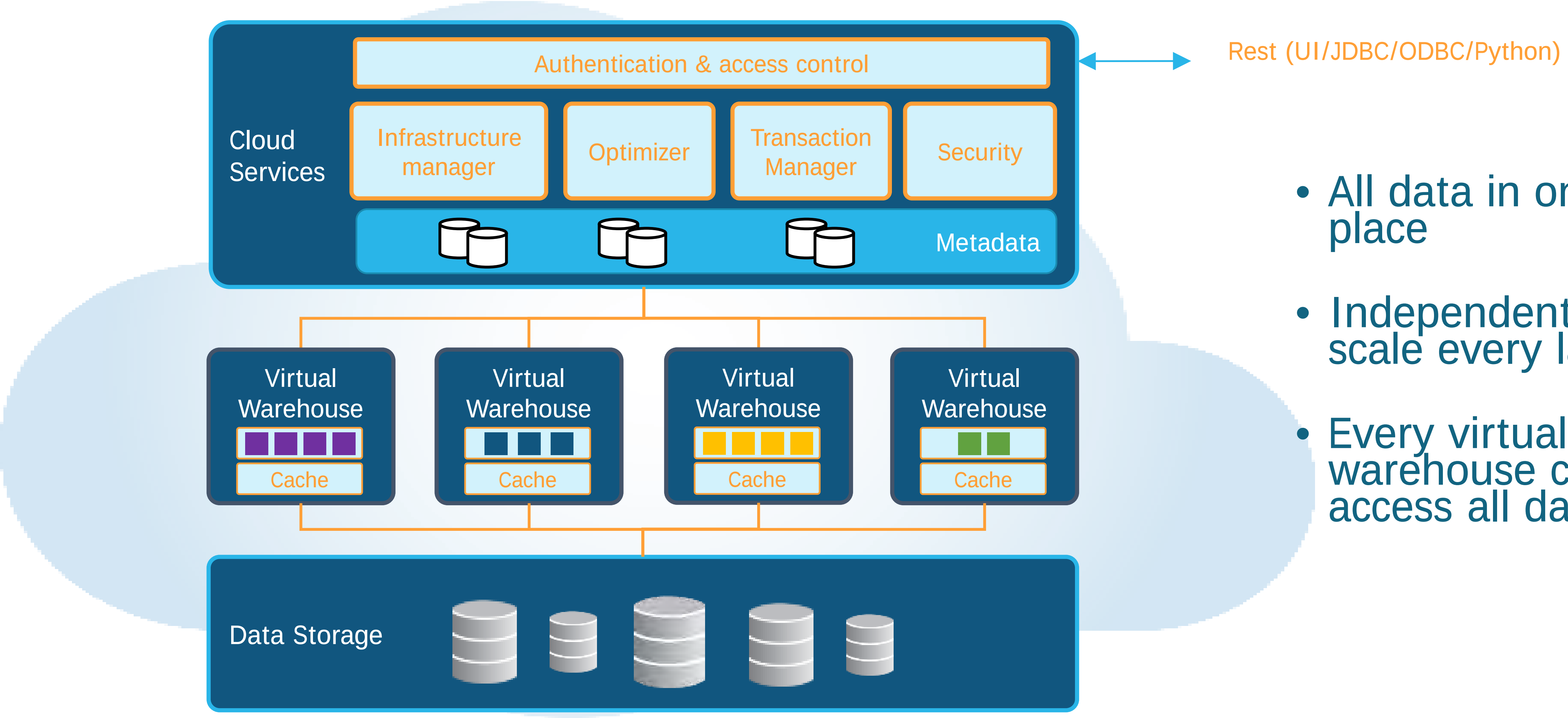
Snowtrail: Testing with Production Queries on a Cloud Database

Jiaqi Yan, Qiuye Jin, Shrainik Jain, Stratis D. Viglas, Allison Lee

Snowflake Database

- The Snowflake Elastic Data Warehouse, or “Snowflake”
  - Analytics database built for the cloud
  - Multi-tenant, transactional, secure, highly scalable, elastic
  - Implemented from scratch (no Hadoop, Postgres, etc.)
  - SQL
- Currently runs on AWS and Azure
- Serves tens of millions of queries per day over hundreds petabytes of data
- 1000+ active customers, growing fast

Multi-cluster Shared-data Architecture



- All data in one place
- Independently scale every layer
- Every virtual warehouse can access all data

Motivation

- Challenges
  - Highly available service, no downtime allowed
  - Fast (weekly) online upgrade process
  - Huge size and variation in customer workloads, hard to exhaustively test

- Opportunities
  - Detailed information of every customer query
  - Multi-tenant capability – easy and secure access from privileged role
  - Resource isolation and elasticity – replay queries with no impact on production workloads

- Our Solution: Snowtrail

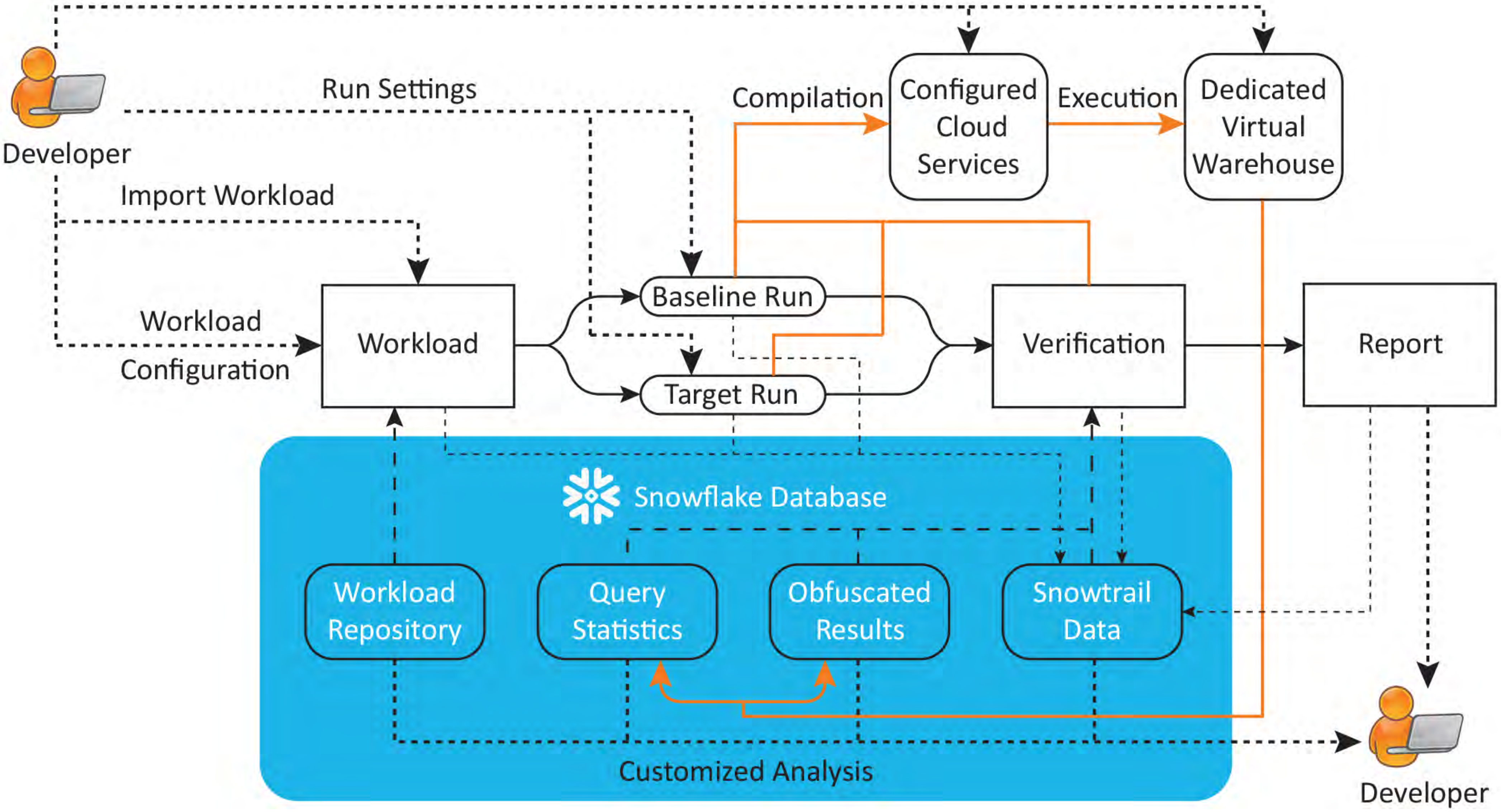
Use Case

- Release Testing
  - Integrated into pre-release pipeline
  - Effective coverage of customer workloads

- Feature Development
  - Incremental Feature Development
  - Immediate Understanding of exact impact

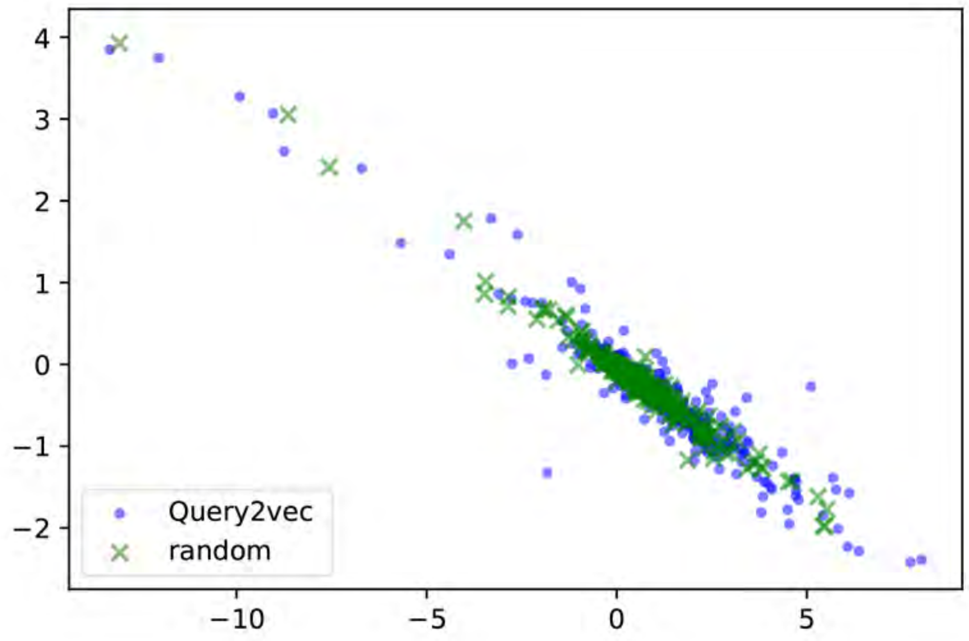
- Workload Runner
  - Cache warming before workload migration
  - Stress Testing
  - Capture and Replay workload for POC

Workflow


Developer
Run Settings
Import Workload
Workload
Configuration
Workload
Baseline Run
Target Run
Compilation
Configured
Cloud
Services
Execution
Dedicated
Virtual
Warehouse
Verification
Report
Snowflake Database
Workload
Repository
Query
Statistics
Obfuscated
Results
Snowtrail
Data
Customized Analysis
Developer

Workloads

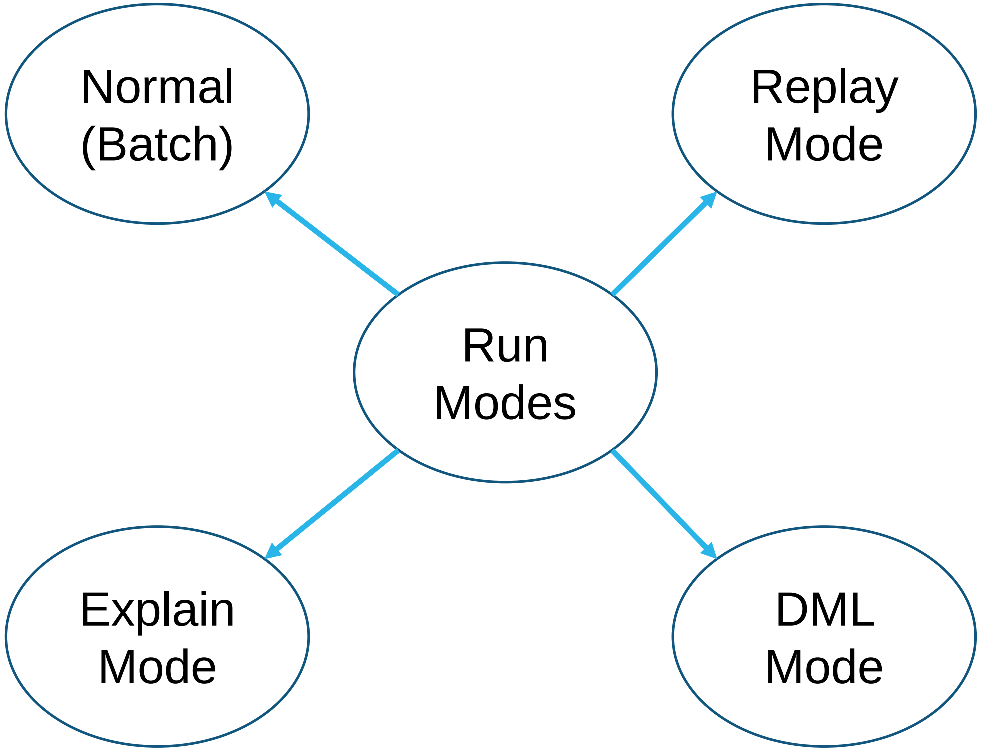
- Sets of queries with their configurations
- Imported workloads
- Workload selection
  - Tens of millions of queries / day – needs sampling
  - Heuristics-based filtering
  - Query2Vec[1]
  - Integration with Usage Tracking



[1] Query2Vec: An Evaluation of NLP Techniques for Generalized Workload Analytics, arXiv:1801.05613

Runs

- Mechanisms:
  - Result obfuscation
  - Query redirection
  - Query compilation context
  - Time travel
  - …
- Configurations:
  - Query compilation context
  - Concurrency
  - Target cloud services
  - Amount of compute resource to use
  - Parameter settings
  - …


Run Modes

Normal (Batch)

Replay Mode

Explain Mode

DML Mode

Analysis

- Look for:
  - New errors / crashes / incidents
  - Wrong results
  - Performance Regression

- Queries can be skipped due to:
  - Change in schema (e.g. dropped tables)
  - Non-deterministic queries are skipped for result comparison
  - False Positives in performance comparison

Verification

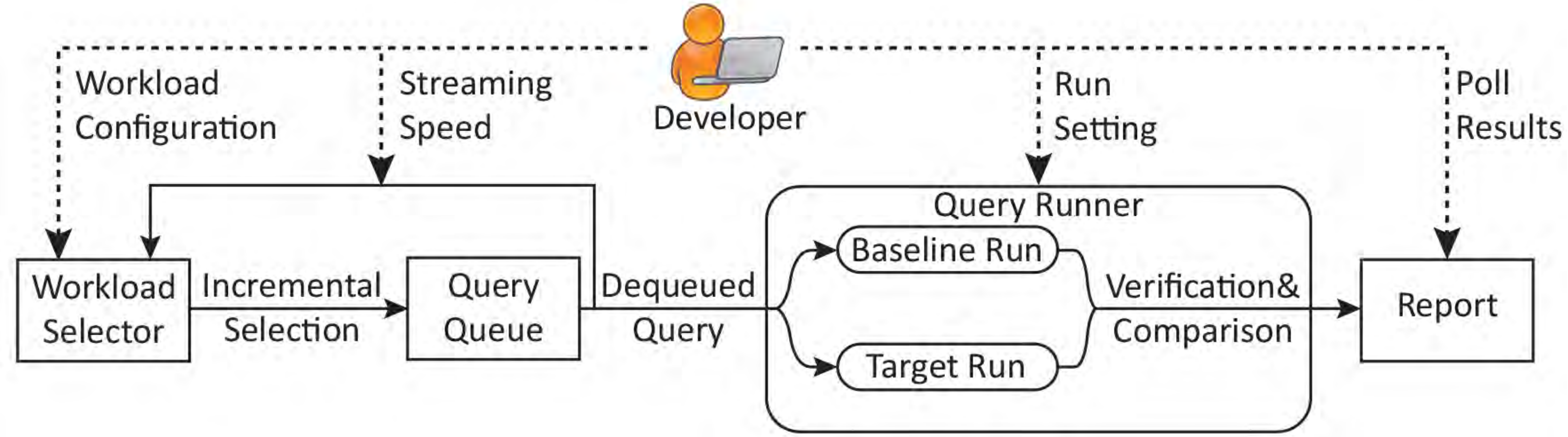
- False positives due to:
  - Cache state
  - Network latency
  - Concurrency
  - …

- Verification runs:
  - Replay regressed queries on the same cloud configuration with isolated resources multiple times

- Result analysis
  - Generate report after verification runs
  - All results stored in Snowflake
  - Run data available in a separate SQL Schema

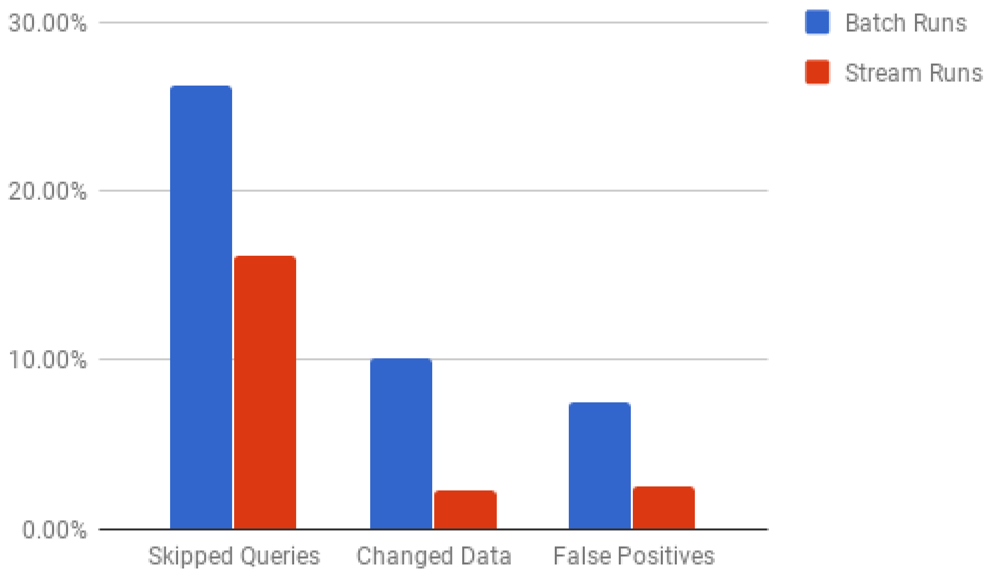
- High false positive rates a major problem

Streaming Mode

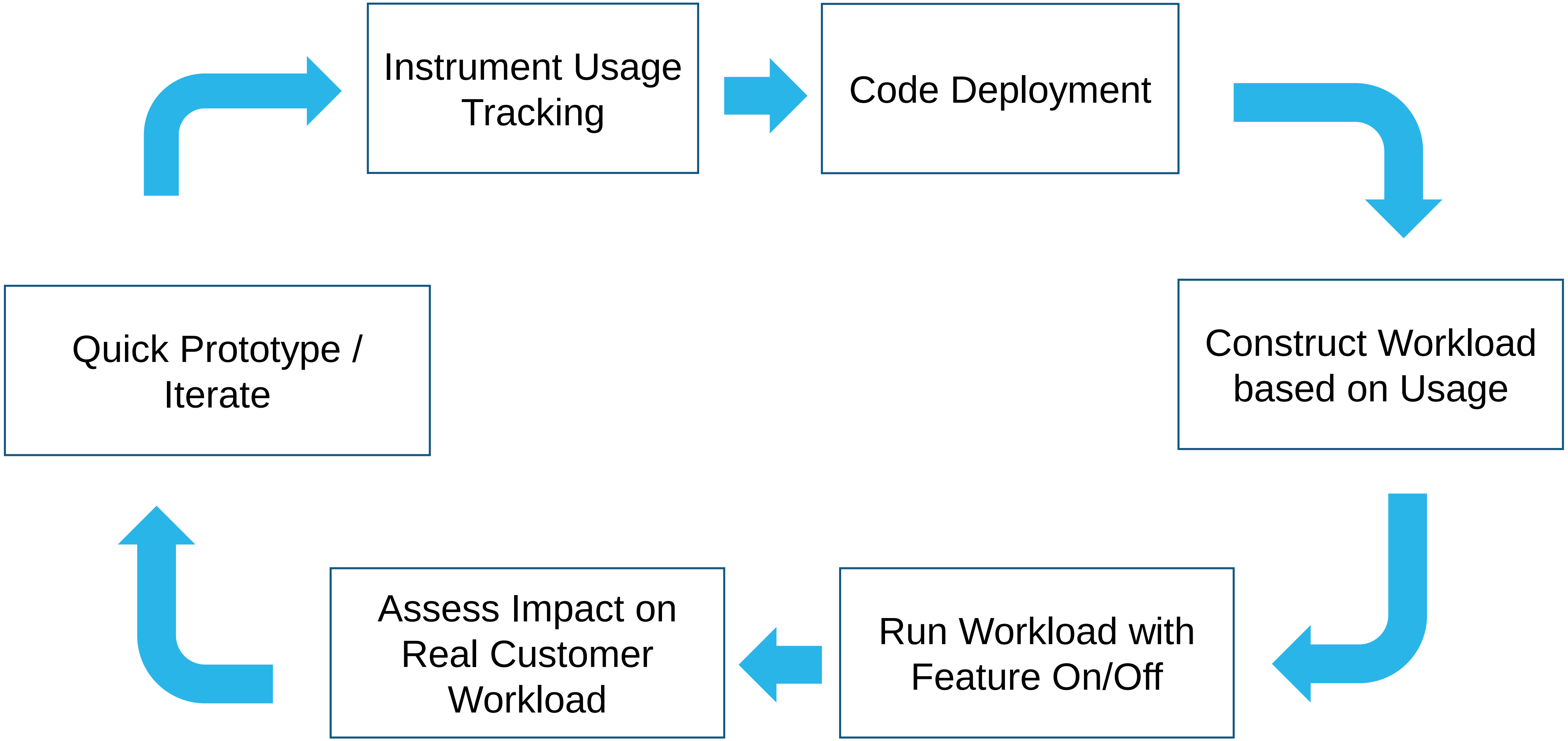


Stream Runs

- Enable continuous runs of arbitrary duration
- Picks up the latest customer queries
- Avoid falling out of time travel retention
- Fewer data and schema changes
- Snapshot reports available



Feature Development Workflow



Instrument Usage Tracking

Code Deployment

Construct Workload based on Usage

Run Workload with Feature On/Off

Assess Impact on Real Customer Workload

Quick Prototype / Iterate

Lessons Learned

- Snowtrail has:
  - Greatly improved release stability
  - Changed how we develop new features
  - Made debugging production queries much easier

- On the other hand:
  - Workload selection is hard
  - Impossible to catch every issue pre-release
  - Complex queries could be expensive to run
  - Lots of non-deterministic queries leads to missed opportunities

Questions?

Thank You

snowflake | YOUR DATA, NO LIMITS