



Escape Vendor Lock-in

Streaming Analytics with Apache Polaris(Incubating)

bit.ly/polaris-k8s-modern-catalog

Kamesh Sampath

Lead Developer Advocate

@Snowflake



-- The Tech Story: SQL Edition

```
CREATE VIEW tech_innovator AS
SELECT
  'Kamesh Sampath' as innovator,
  'Veteran Tech Innovator' as identity,
  '20+ years in the trenches' as experience,
  ARRAY ['Author', 'Developer Advocate'] as roles,
  'Data and Cloud Technologies' as domain,
  'Empowering developers' as mission;
```

The Human Story: Python Edition

```
def share_mission ():
    return """
    Tech innovator with 20+ years in the trenches.
    As an author and developer advocate, I'm on a quest
    to demystify data and cloud technologies.
    Mission: Empowering developers to push boundaries
    and create the impossible.
    """
```



order user order
user emp user
emp order emp

"Data Streams" is a continuous flow of data from variety of data sources

```
{
  Joe joined Acme
  as a developer on
  25/10/2023
}
```

```
{
  Order for two pizzas
  placed at 12:00PM
}
```

"Event" is a digital record of an action at a of time.

Event Driven Architecture (EDA)

A software architectural style to build applications using events



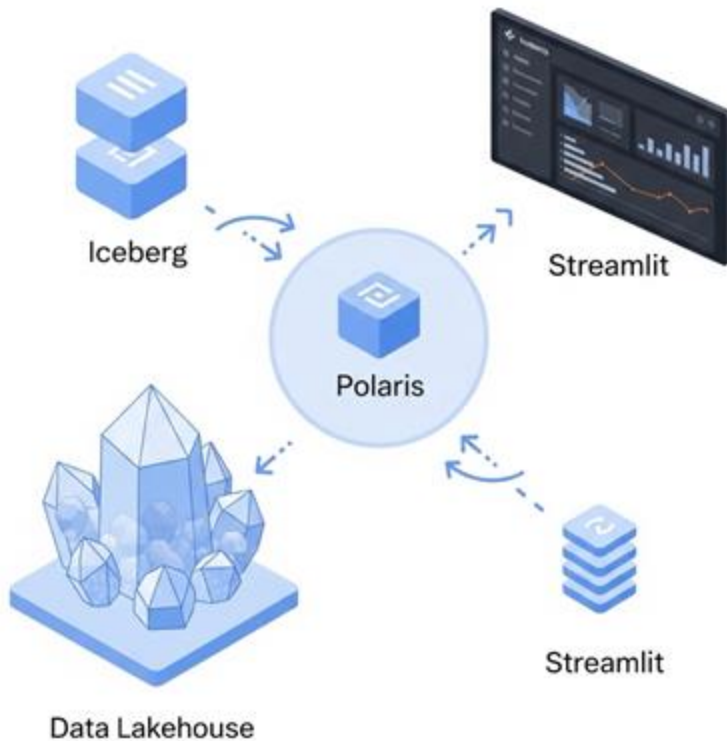
Event Producer



Event Consumer



Balloon Game App



Apache Kafka



Key Capabilities

- **Stream:** Publish and subscribe to streams of records
- **Store:** Store streams of data with fault-tolerance
- **Ship:** Process streams of records as they occur

Why Kafka Matters ?

- Handles trillions of events per day
- Low-latency (milliseconds)
- Scalable, fault-tolerant architecture
- Decouples data streams and systems
- Enables real-time data integration across the enterprise

Use Cases:

- Real-time analytics
- Event-driven architectures
- Microservices communication
- Log aggregation
- Stream processing





RisingWave – Quick Introduction

What is RisingWave?

- Open sourced in April 2022 under Apache 2 License.
- Purpose-built for Stream processing workloads



Key Stats:

- ~8K GitHub stars
- ~200 GitHub contributors
- ~2.5K Slack members
- ~300K K8s deployments

Used and loved by



Team Background



What s Apache ICEBERG™?



- **Open Table Format**
- **Processing Engine Agnostic**
- **Advanced Table Features**
- **Performance Optimized**
- **Enterprise-Ready Governance**



What is a Data Catalog?



- Centralized inventory of an organization's data assets.
- Essential for data discovery and understanding.
- Improves data governance and compliance efforts.
- Enhances collaboration among data teams.

WHAT IS Apache Polaris(Incubating)?



Incubating under Apache Software Foundation

- Platform Independent
- Programming Language Agnostic
- Vendor Neutral
- Portable Metadata
- Enterprise-Grade Catalog

The modern way to connect data applications to your Iceberg tables



What is Streamlit



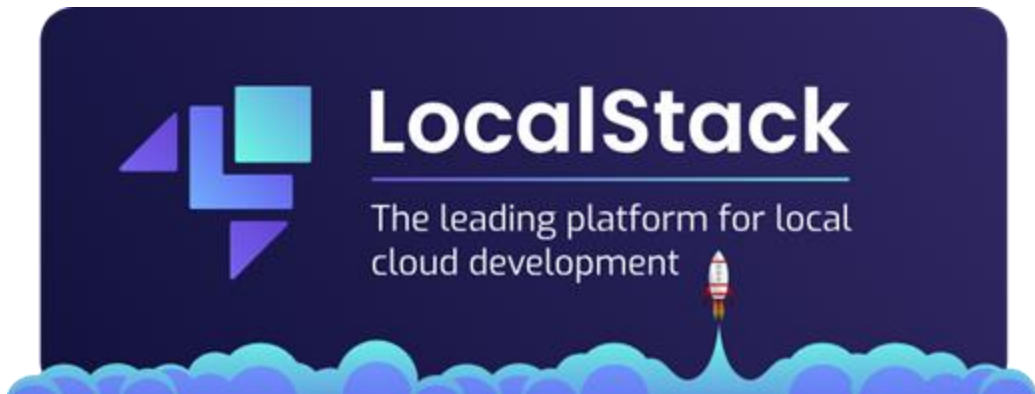
Streamlit

- An OSS Web framework in Python
- Turns data into shareable web apps in mins
- No front-end experience is required
- Focus on data and model



What is Localstack?

A fully functional local cloud stack - *Develop your AWS cloud apps locally!*



- Enables a highly efficient dev and test loop for cloud apps
- Ships as a Docker image, easy to install and start up
- Support for ~100 APIs (and growing):
 - compute (e.g., Lambda, ECS, EKS)
 - various databases (e.g., DynamoDB, RDS)
 - messaging (e.g., SQS, Kinesis, MSK)
 - some sophisticated/exotic APIs (e.g., QLDB, Athena, Glue)
- Advanced collaboration features and CI integrations
 - redefining the way cloud apps are developed across the lifecycle!



Why I Built This ?

- Personal Pain
- Democratization Goal
- Learning Journey
- Focus Shift
- Evaluation Need
- Developer Friendliness
- Cloud-Native

Enterprise Tools vs Ease of Adoption

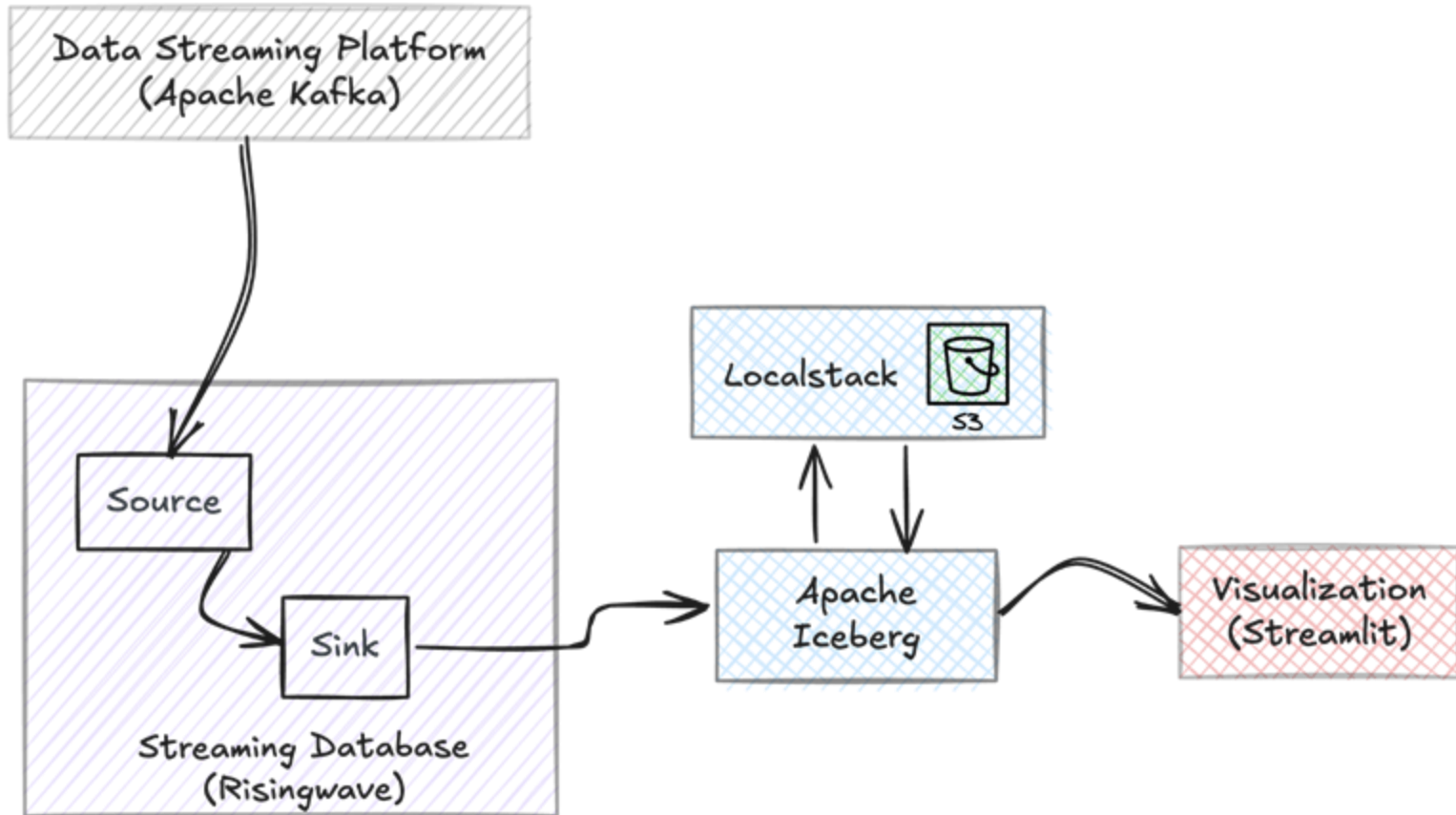
- Bloatware Problem
- Developer Experience Gap
- Laptop-Friendly Development
- Cloud-Native Integration

Enterprise Tools vs Ease of Adoption

- Time-to-Value Divide
- Expertise Barrier
- Testing vs Production Reality



Wiring all together



Building analytics with Streaming Data

- Applying DRY
 - One Source → Multiple Sink
- KISS
 - Just SQL
- Even better
 - If it is Cloud Native
 - If I can build time-series

DEMO



Core Benefits

- Automation
- DevOps Practice
- Deployment Pipeline
- Open Source
- Developer First
- Cloud Native
- Best Practices

Resources



bit.ly/polaris-k8s-modern-catalog

Thank you!

