Stream Me Up, Scotty

Transitioning to the cloud using a streaming data platform

Strata Hadoop World
San Jose, CA
March 15, 2017
About us

Gwen Shapira

Bob Lehmann
We’ll Talk About

● Everyone is moving to the cloud
● Common challenges
● Transition design patterns
● How streaming platform fits in
● Monsanto’s data challenges
● Monsanto’s streaming architecture / use cases
● Security
● Lessons learned
● Future plans
How Amazon thinks you will move to cloud
You have a lot to worry about

- Do I migrate to cloud-native databases?
- Will my apps survive random cloud failures?
- Do I need a failure injection test framework?
- Dynamic configuration of hostnames
- Do I break my monolith into microservices?
- Do I want to migrate to more than one cloud?
- Can I migrate back?
At first, this is no big deal....
6 MONTH LATER...
Are you kidding?

- This is expensive
- This is a maintenance nightmare
- We may need more than one region!
- We may need more than one cloud!
We’ve done this before...

This...

To this...
There is a better way
Bridge to Cloud Pattern
Key Components

1. Distributed Log
2. Big Reliable Buffer
3. Great Replication
4. Connect Anywhere
5. Management & Monitoring
Deeper look at bi-directional replication
PLANNING

No challenge is too great if you plan ahead. And have pointy ears.

Why is this awesome?

- Proven architecture
- Non-stop low-latency
- One throat to choke
- Cost savings
But wait - there is more!

- Future-proof architecture
  - With Connect - get the data everywhere.
  - Try cloud services without fear of lock-in
  - “Kafka is our escape valve”
- Multi-zone availability
- Multi-Region architecture
- Multi-Cloud architectures
- Microservices ready
- Jump to stream processing!
Let’s look at how this worked for...
Monsanto
A sustainable agriculture company

- Bringing a broad range of solutions to help nourish our growing world
- Headquartered in Saint Louis, Missouri
- >20,000 employees in 66 countries
- A global company with >50% employees based outside of the United States
- One of the 25 World’s Best Multinational Workplaces by Great Place to Work Institute

“ produc e with more judicious use of limited natural resources.

improve the lives of the world’s farmers.

Increase production to meet needs of a growing population.

“We succeed when farmers succeed.”
Hugh Grant, Monsanto CEO

FORTUNE WORLD’S MOST ADMIRE D COMPANIES

DiversityInc Top 50
World’s Most Diverse

Science 2015 Top Employer

DIGITAL EDGE 25 AWARDS

CR’s 100 Best Corporate Citizens

BEST PLACES TO WORK 2015 for LGBT Equality

50 SMARTEST COMPANIES
The Monsanto Corn “Galaxy”
To Boldly Go Where No Man (or Woman) Has Gone Before...

CAPTAIN’S LOG STARTDATE 41153.7 (early 2015)

Mission: Develop an enterprise information architecture covering BI to machine learning and everything in between. Maximize use of cloud based assets.
In What Parallel Universe Are You Living?
Existing Challenges

- Data sprawl
- Data consistency
- Data discovery
- Data latency
The Cloud - New Challenges

- Increased data sprawl (microservice’s dirty little secret)
- Can’t forklift applications overnight
- Cloud apps need on-prem data
- On-prem apps need cloud data
The “Aha” Moment

The Log: What every software engineer should know about real-time data's unifying abstraction

LinkedIn Engineering Blog post by Jay Kreps
Dec. 6th, 2013
Let’s Clean Up This Mess!

Courtesy of Jay Kreps
Replication Is Your Friend

- Uncontrolled, unsynchronized replication IS BAD
- Controlled, synchronized replication IS NECESSARY AND BENEFICIAL...especially in distributed environments
The Enterprise DataHub

EDH Plan

- Create Kafka clusters on prem and in AWS
- Establish cross datacenter connection
- Provide replication between the clusters
- Use AVRO schemas
- Only allow apps to interact with the local cluster

VPN? DIRECT CONNECT?

MIRROMAKER

IMPORTANT!
USE SCHEMA REGISTRY
Use Cases And Architecture Evolution
Use Case - Customer 360

Diagram showing various components and their interactions, including:
- ERP System
- Schema Registry
- Zookeeper
- Kafka
- Cloud Foundry
- EC2 Container Service
- REST Proxy

Connections between these components illustrate data flow and integration.
Use Case - Data Warehouse Replication

Oracle
Informatica Data Replication
Teradata Staging
ETL
Teradata Final

ERP System

Schema Registry
Zookeeper
Kafka Connect Cluster
JDBC Connector
Table 1
Table 2
Table 3
Table 4

Kafka
Replikant

Schema Registry
Zookeeper
Kafka
Replikant
Table 1
Table 2
Table 3
Table 4

Kafka Connect Cluster
JDBC Connector

Postgres
Apps

Confluent 3.1.2 / Kafka 0.10.1.1
Use Case - Data Warehouse Migration

Turn this feed off once cloud solution is complete.
Cross Datacenter Replication - The Killer App!

Drives adoption of the platform despite these common “concerns”:

- Our data will never be used by anybody else
- We don’t need our data in near real time
- Our volumes are too low
- Our volumes are too high
- Do we really have to use schemas?
Enterprise DataHub

Simplified AWS Architecture

Note: CIDR blocks for all VPCs and on-prem network must be non-overlapping
Security

Native Clients

- SSL certs used for authentication AND authorization
- Principal is defined by the cert DN
- ACLs are defined based on DN

CN=YourProject, OU=YourOrg, O=YourCompany, L=YourLocation, ST=YourState, C=YourCountry

CN and OU are the only attributes defined by clients
Security

REST Clients

- API endpoint defined in Network Director for each topic
- REST client authenticates with SSO platform and then calls API endpoint
- Network Director calls REST Proxy as privileged user
- Access to topics is controlled by restricting access to API endpoint
Security Notes

- Some Kafka tools don’t work with SSL and ACLs
- SSL client certs have to be added to truststore on each broker - requires brokers to be bounced. Patched broker code to load certs dynamically.
- Long term roadmap
  - LDAP authentication and role based authorization for native clients.
  - REST Proxy authorization based on OAuth entitlements
Lessons Learned

- Partition By Default
- Use “new” consumer - required for SSL anyway
- Lock down Zookeeper
- Be prepared for AWS instance termination!
- Use Associated VPCs in AWS
- Use LVM For disks
Future - Make Ingress/Egress Part Of The Platform
Future - Multi-Cloud Integration
THE END