



# **Flower:** A Streaming Data Analytics Flow Manager

Alireza Khoshkbarforoushha, ANU and CSIRO

Supervisory Panel: A/Prof. Rajiv Ranjan, Newcastle University, UK

Prof. Alistair Rendell, ANU



## Availability of Data-Intensive Systems



## Availability of Data-Intensive Systems

## Affordability of Infrastructure Resources





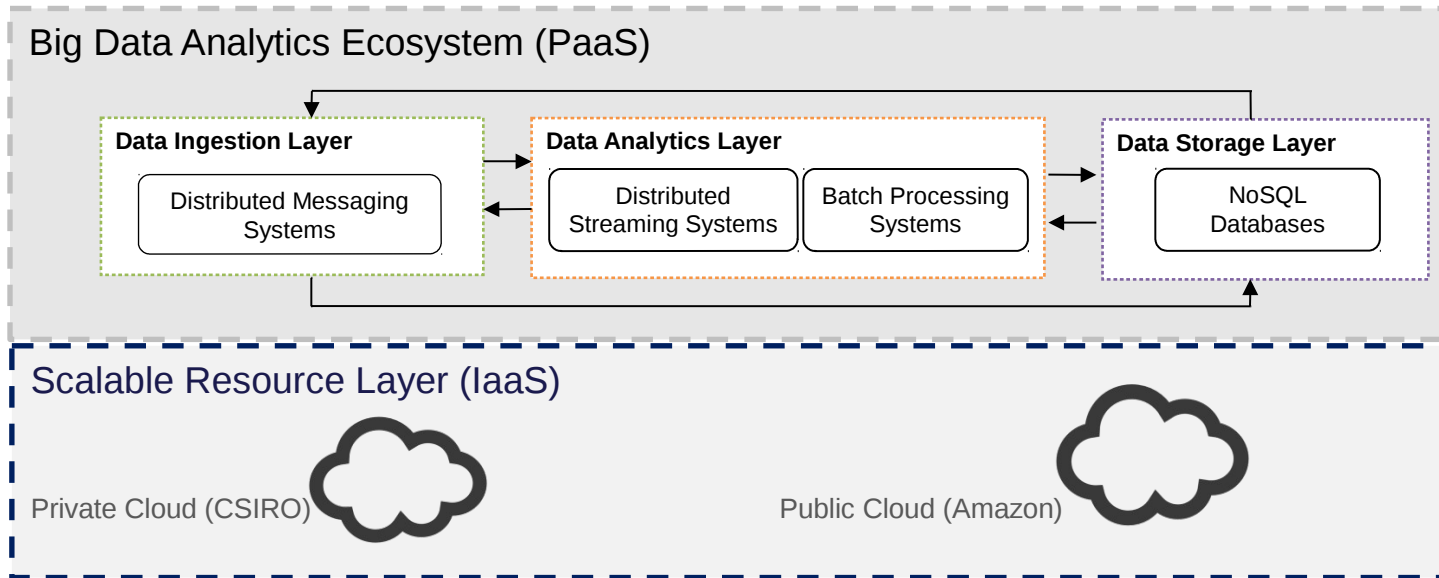
## Availability of Data-Intensive Systems

- Click Stream Analytics
- Emergency Situation Awareness (ESA)
- Sentiment Analysis
- Trending Topics for Event Detection (e.g. flood, flu, etc.)

## Affordability of Infrastructure Resources

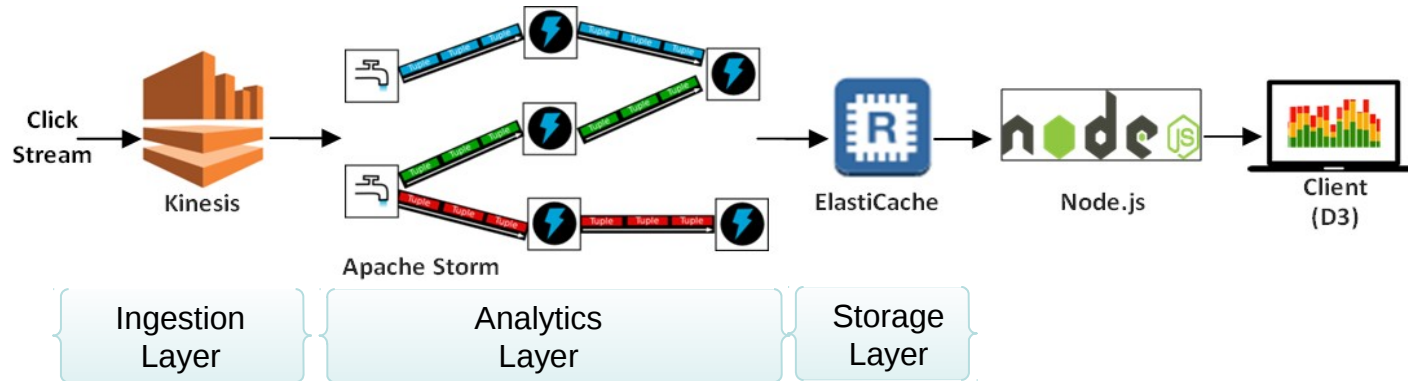


# Large Scale Analytics Flow Ecosystem



# Example

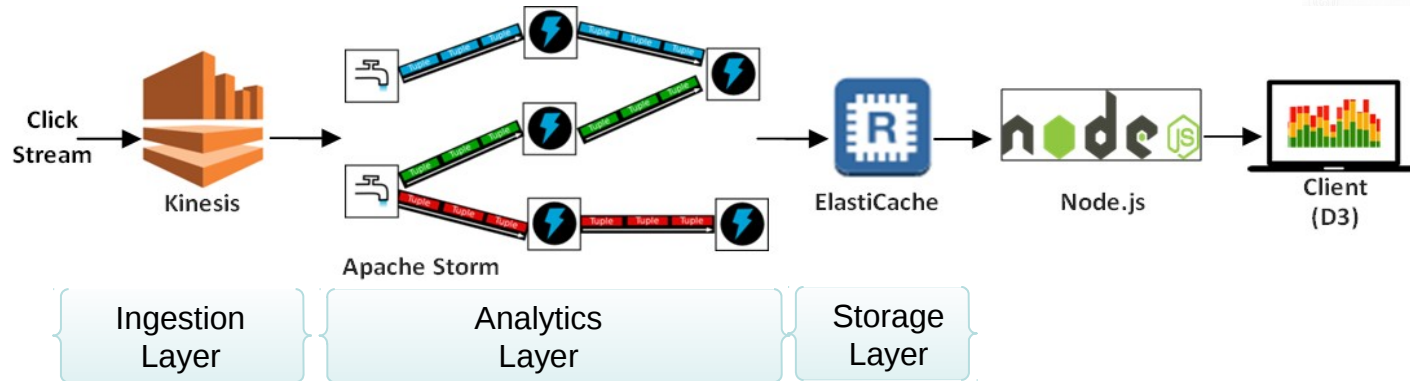
- Click Stream Analytics



# Problem?

- How to:

**Evaluate** performance measures across the Flow

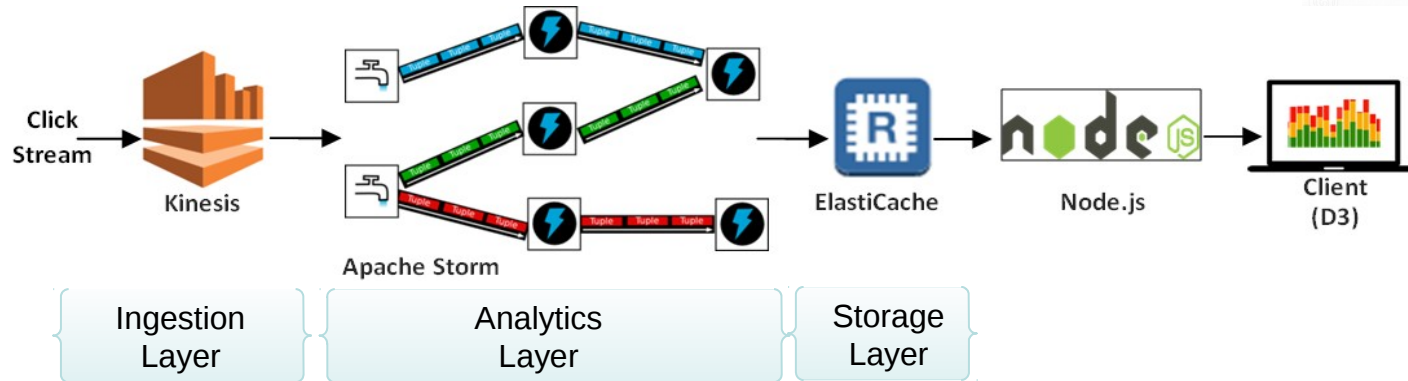


# Problem?

- How to:

**Evaluate** performance measures across the Flow

**Locate** abnormal behaviors





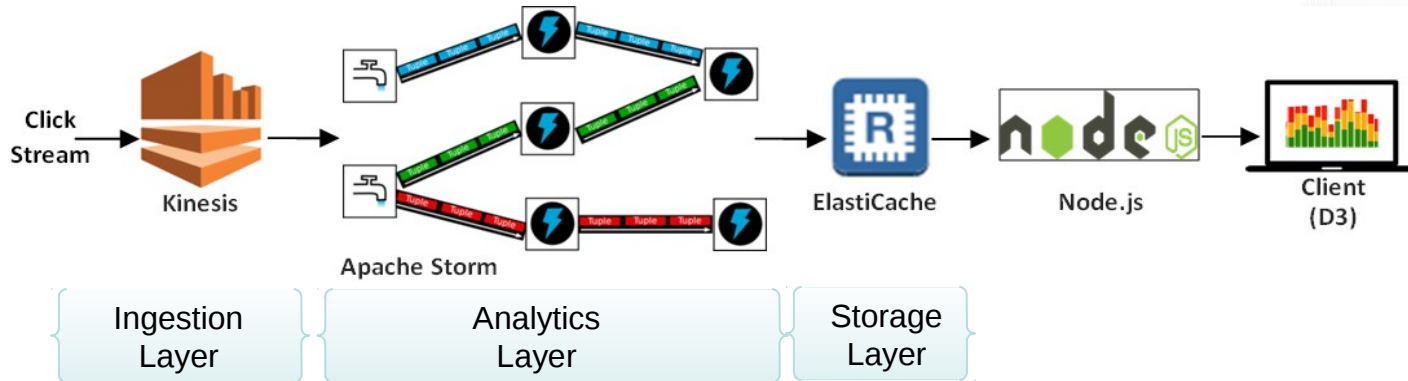
# Problem?

- How to:

**Evaluate** performance measures across the Flow

**Locate** abnormal behaviors

**Scale** if data arrival rates increase 10X



# Problem?

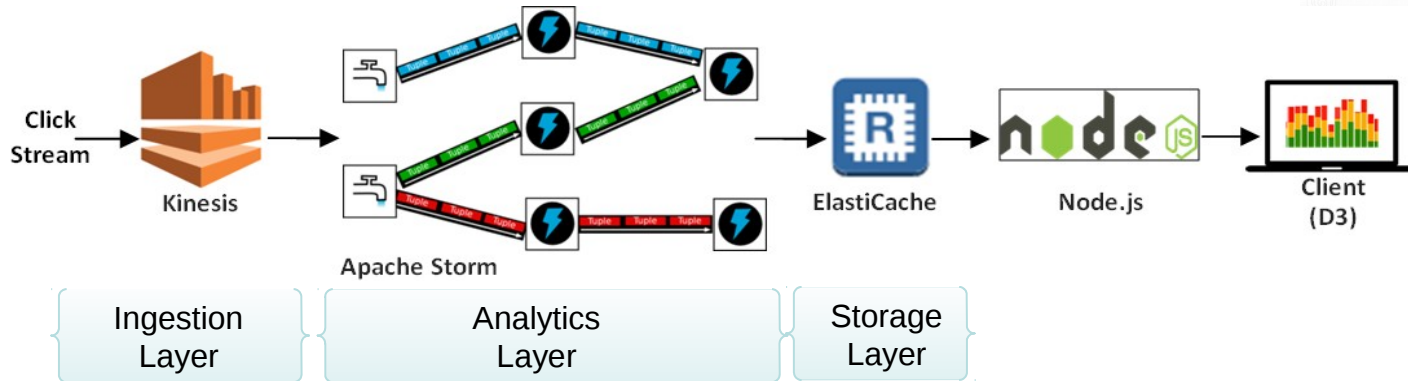
- How to:

**Evaluate** performance measures across the Flow

**Locate** abnormal behaviors

**Scale** if data arrival rates increase 10X

**Manage** monetary costs of the Flow





# Goal:

Building a system for seamless management of the analytics flows.



# **FLOWER** core functionalities...



# 1. All-in-One-Place **Visualization**

Cross-Platform Monitoring of Performance Measures

# 1. All-in-One-Place Visualization

## Cross-Platform Monitoring of Performance Measures

Features:

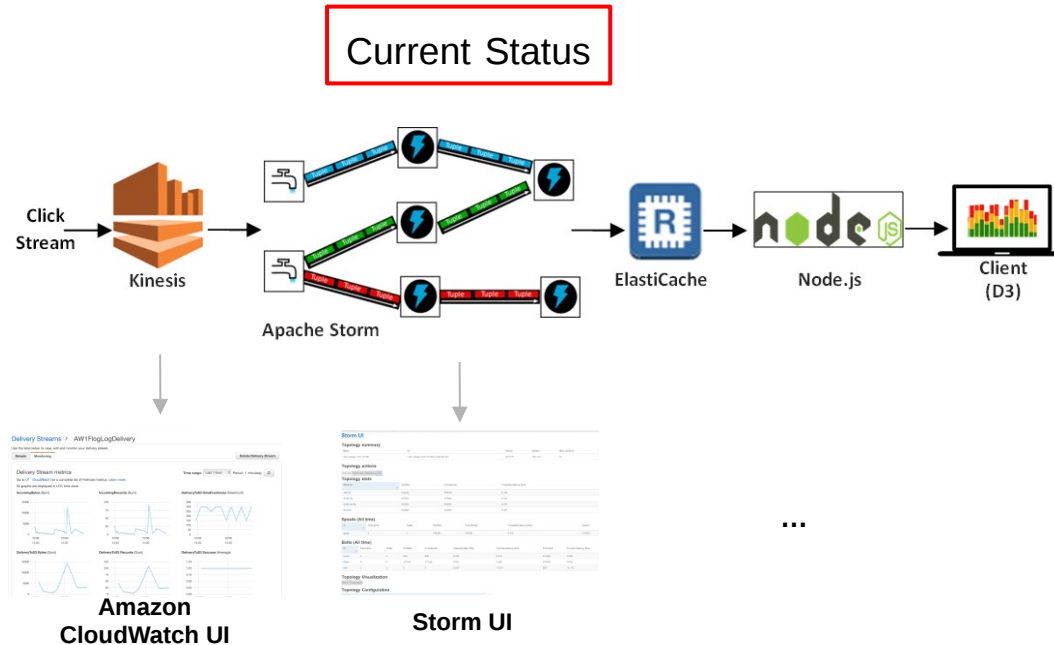
### 1.1. Unified monitoring dashboard

- Holistic performance measure generation
- Drill-down analysis

### 1.2. Error and Warning Summarization (distributed in many machines)

Benefits:

- Better visibility to performance issues
- Faster performance inspection





## 2. **Elasticity** Management of the Flow

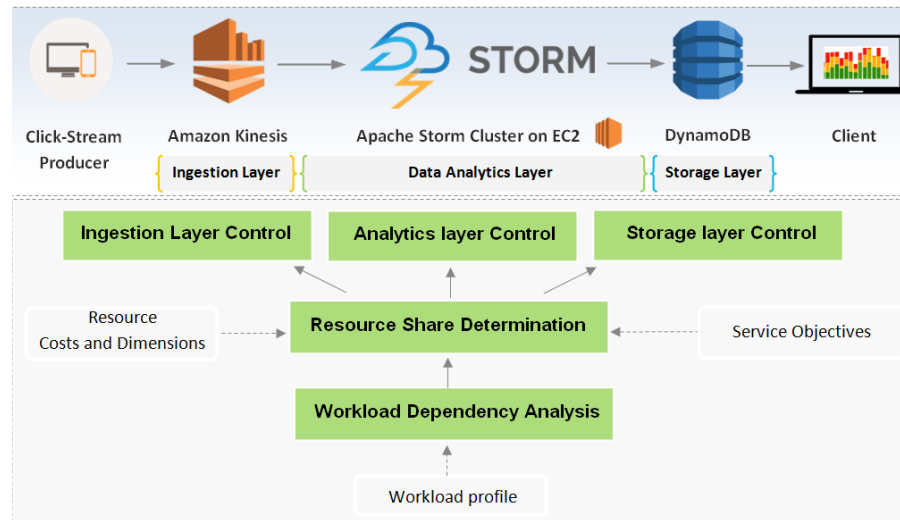
## 2. Elasticity Management of the Flow

### Features:

- Dynamic resource provisioning using control theory
- Cost-based optimization across the Flow

### Benefits:

- Automated control of heterogeneous resources





# 3. Resource Cost Analysis of the Flows

Current Status

Features:

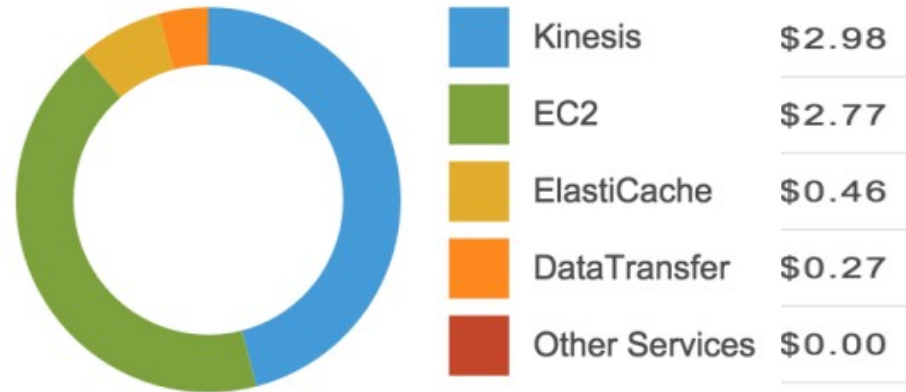
3.1. Cost management dashboard

- Resource costs per analytics flow


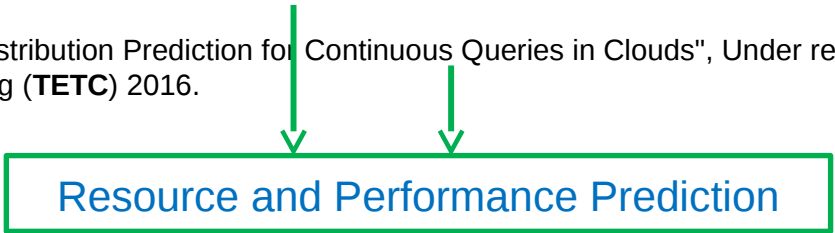

3.2. Alternative deployments costs w.r.t pricing models (e.g. on-demand, spot, reserved)

Benefits:

- Detailed cost management of the applications



## Flower is mainly built upon our following studies:

- A. Khoshkbarforoushha, et al., "Dimensions for Evaluating Cloud Resource Orchestration Frameworks", **IEEE Computer**. (In Press, 2016)  

- A. Khoshkbarforoushha and R. Ranjan; "Resource Usage Distribution Prediction for Continuous Queries in Clouds", 7th ACM/SPEC International Conference on Performance Engineering (**ICPE**) 2016, Delft, the Netherlands.
- A. Khoshkbarforoushha, et al.; "Resource Usage Distribution Prediction for Continuous Queries in Clouds", Under review in IEEE Transactions on Emerging Topics in Computing (**TETC**) 2016.  

- A. Khoshkbarforoushha, et al.; "Elasticity Management of Streaming Data Analytics Flow on Clouds", Under review in Journal of Computer and System Sciences (**JCSS**), Special Issue on Theoretical and Algorithmic Foundation for Big Data, 2016.  


# Wrap-up

- Management of the Big Data Analytics Flows are problematic:
  - Heterogeneous workloads
  - Diversity of platforms
  - Diversity of Cloud services
- Our Solution: **Flower**, a system that:
  - Gathers and shows **all relevant data** about the performance in one place
  - Automatically **controls** the resource requirements of the flows
  - Provides necessary grounds for **cost analysis** of the applications



**Thank You.**

