Big Data Lab, Colorado State University

A REAL-TIME ANOMALIES DETECTION SYSTEM BASED ON STREAMING TECHNOLOGY

Khushboo Gupta

OVERVIEW

- Background
- Introduction
- System Architecture
- Results
- Conclusions

BACKGROUND

- Discovering periodical changes and temporal trends of packets count.
- **Apache Storm** provides a distributed and fault-tolerant real-time computation method, makes it easy to process unbounded streams of data.

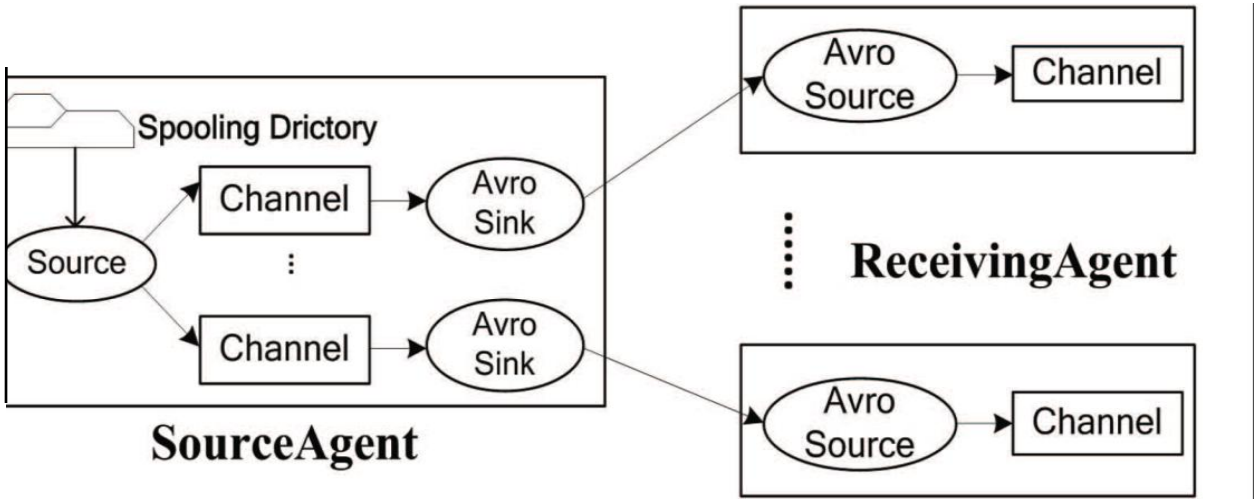
INTRODUCTION

- Implement a distributed streaming computing system
- Monitor the mutation of flow data
- Locate the source of anomalies
- Find the specific abnormal IP addresses

SYSTEM ARCHITECTURE

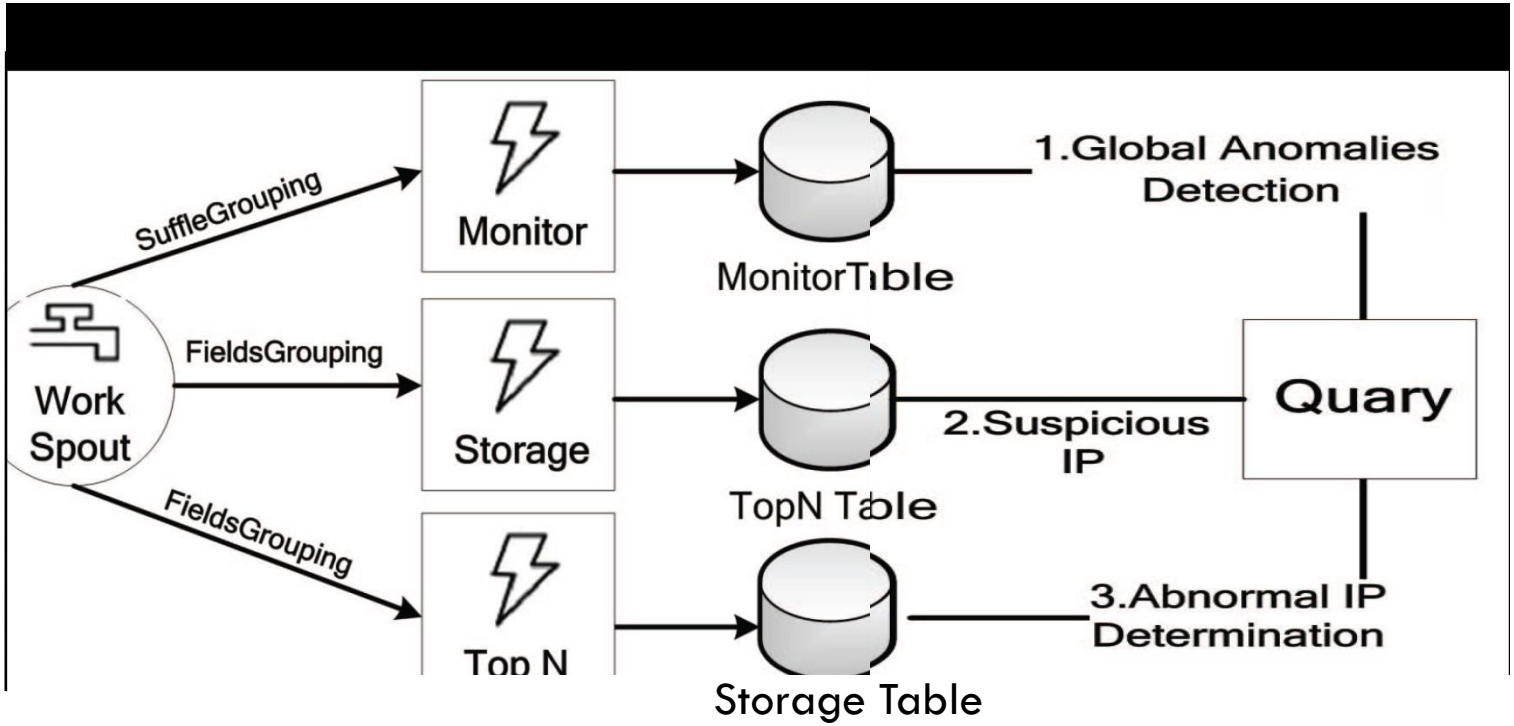
- Flexible integration of Storm and distributed log collection system Flume, working as collector of flow data

1. Data Collection



SYSTEM ARCHITECTURE

2. Data Processing



SYSTEM ARCHITECTURE

- TABLE I. HBASE MONITOR TABLE STRUCTURE

Row Key	Column: Qualifier
each time unit (eg.1392301200)	0: no abnormal 1: uplink abnormal 2: downlink abnormal 3: both way abnormal

SYSTEM ARCHITECTURE

- TABLE II. HBASE TOP N TABLE STRUCTURE

Row Key	Column: Qualifier
each time unit (eg.1392301200)	N IP addresses: IP1,IP2,...IPn-1,IPn

SYSTEM ARCHITECTURE

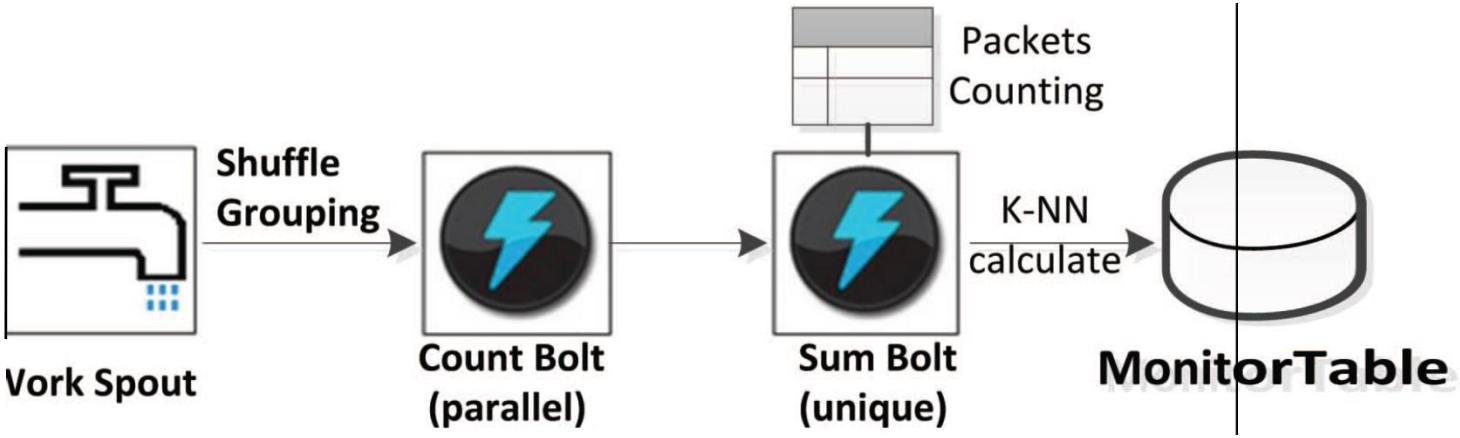
- TABLE III. HBASE STORAGE TABLE STRUCTURE

Row Key	Column: Qualifier
IP+","+each time unit (eg.123.456.78.9,1392301200)	Up/down PKTs count

SYSTEM ARCHITECTURE

3. Real time anomalies detection algorithm

- K-NN algorithm
- Two stages- count bolt and sum bolt



RESULTS

1) Real-time TopN Computing

- *Accuracy* = *M/N*

  M is the intersection of "*TopN table*" *and authentic top N IP* set.

Time/ Accuracy	Top 10	Top 25	Top 50	Top 100
22:05	100%	100%	100%	96%
22:15	100%	100%	100%	100%
22:25	100%	100%	100%	97%
22:35	100%	100%	100%	96%
22:45	100%	100%	100%	98%
22:55	100%	100%	100%	98%

RESULTS

2) Real-time anomalies detection

- Anomalies detection method that makes k-NN algorithm embedded in Storm topology works well.

3) Performance and Scalability analyze

- memory did not become the bottleneck of performance.
- CPU properties dominate performance

RESULTS

4) Scalability Evaluation

- Linear scalability

Num of Nodes	Maximum processing records (*10000 per min)	Maximum processing records Per Node (*10000 per min)
2	80	40
3	118	39.3
4	152	38
5	182	36.4

CONCLUSIONS

- Proposed system achieves accurate real-time anomalies detection from mass flow data in a scalable way.

QUESTIONS