Real Time Analysis of DDoS

-Architecture And Implementation

UTKARSH KAJARIA

MS DATA SCIENCE

UNIVERSITY OF MINNESOTA

OUTLINE

- Requirements
- Technologies
- Scalability: Web Servers
- Scalability: Use Cases
- Architecture
- Analysis
- Demo

REQUIREMENTS

- Ingest server log data from HDFS
- A tool for putting the ingested data on to a message system
- An application that analyzes the logs to identify if IP addresses are part of a DDoS attack
- Store the identified IP addresses for further downstream processing
- Workflow latency of 1-2 minutes

THINKING ABOUT TECHNOLOGIES

- Kafka
  - One of it’s kind: broadcast + message queue
  - High throughput
  - Persists messages with replication
- Spark
  - Versatile: Streaming, ML (in addition to SQL and graph)
  - Integrates well with Hadoop
  - Used extensively hence good support with in the development community
- Spark Streaming + Kafka Integration Guide

PRODUCERS, CONSUMERS, BROKERS

- Producer : Reads from HDFS, puts on Kafka topic

- Consumer : Receives messages, processes, stores.

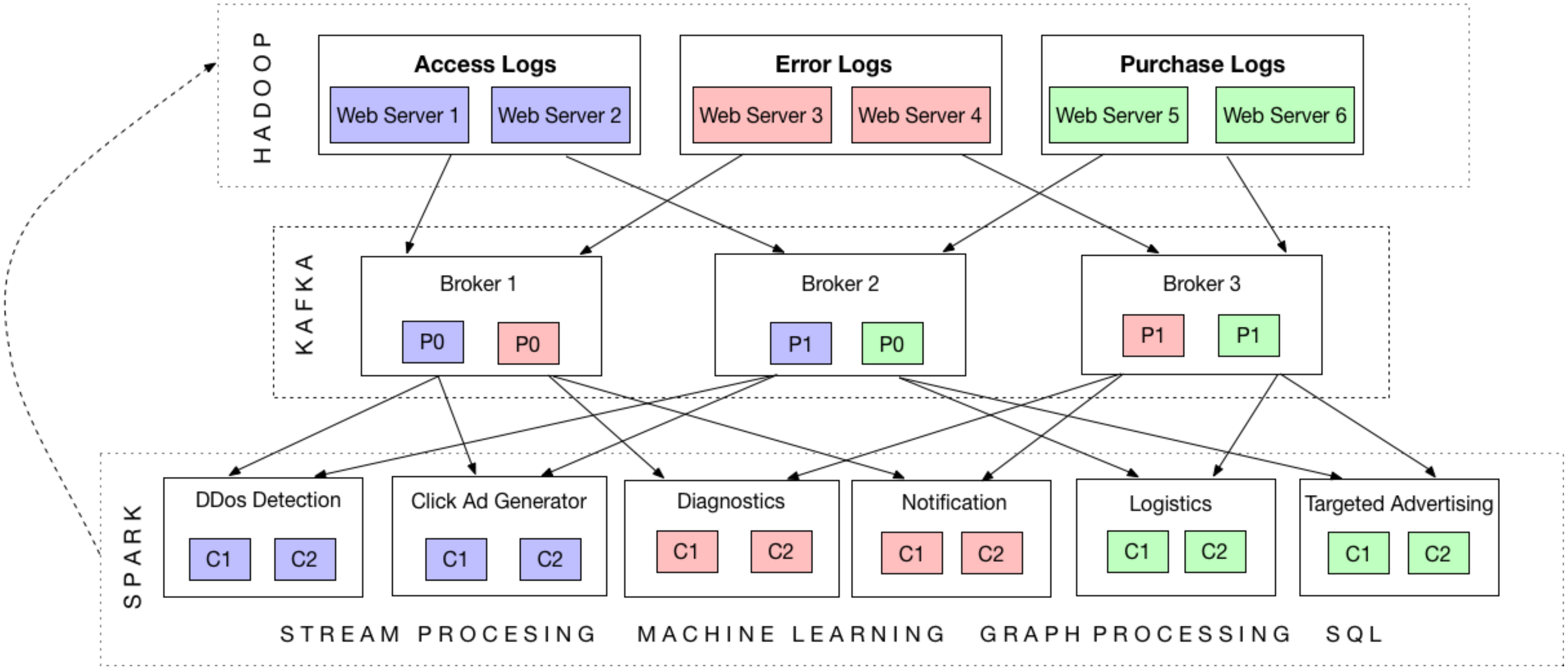
THINKING ABOUT SCALING :SERVERS

- Handling new web server - new producer, new topic or both?
  - New producer
    - Decoupling from other processes already running.
    - Parallelism, rather than bottleneck
    - No single point of failure
  - New Topic
    - Existing consumers have to be restarted with a fresh topic list, or,
    - Will need a new consumer. (Not ideal, only if new use-case)
    - Receiver down – Server stops getting processed

THINKING ABOUT SCALING :USE CASES

- Consumer groups
  - Parallel instances in a group: Load balancing
  - Same topic: Multiple groups for multiple use-cases
- Partition
  - Producer can assign for each record based on a key
  - Distributed evenly and dynamically over consumer instances
  - Message consumed only once by the group

ARCHITECTURE



PROCESSING

Options

- Number of requests/IP address in a time window
- Bytes of data requested is high
- Total requests exceed a threshold
- Unusually high Response time for a request or Http response code 503
- Classification: legitimate or bot
- Anomaly detection using Time Series Analysis

Approach

- Trigger: Total requests
- Identification: Number of requests/IP address

Crunching Numbers

Yelp gets 150m requests a month

~ 58 requests/sec

Our measures

A window of 30 sec

~ 1500 requests triggers a job

~ 60 or request by each IP gets recorded

DEMO

THANK YOU

QUESTIONS?