Cloudberry - Big Data Visualization

Sadeem Alsudais, Qiushi Bai, Chen Li

UC Irvine
BOSS Workshop 2019
Big Data Visualization Tools

Can't Upload Files

File too large. Maximum size: 1.0 GB

graph1.csv
Big Data Visualization Tools
Our solution: Cloudberry

A middleware solution for interactive analytics and visualization on large data
Cloudberry Architecture
Prototype: Twittermap

1.6+ billion records; 2TB; temporal/spatial/textual conditions; Hardware: < $6K
Tutorial Overview

- Twittermap demo
- Cloudberry overview
- Instructions to setup a Cloudberry application on social media visualization
- Under-the-hood details
Cloudberry Tutorial

Time: 11 AM & 2 PM
Location: Santa Monica (3rd level)
Cloudberry - Big Data Visualization

Sadeem Alsudais, Qiushi Bai, Chen Li

UC Irvine
BOSS Workshop 2019
Tutorial Overview

- Twittermap demo
- Cloudberry overview
- Instructions to setup a Cloudberry application on social media visualization
- Under-the-hood details
Twittermap Application

http://cloudberry.ics.uci.edu/apps/twittermap/
Twittermap Settings

- # of tweets: >1.6B (2TB)
- Continuous tweet ingestion
  - 3M tweets / day
- A cluster of 5 Intel NUC machines
  - Intel Core i7
  - 32GB memory
  - Samsung 1TB EVO NVMe SSD
  - < $6K
Cloudberry

A middleware solution for interactive analytics and visualization on large data

http://cloudberry.ics.uci.edu/
Cloudberry Architecture
Cloudberry Architecture
Metadata

1. DDL statement

2. Collect info

3. Save

Cloudberry

MetaData Manager

AsterixDB

Tweets

("hurricane", updated till: May 18)
Cloudberry Architecture
Answering Queries Using Views

Towards Interactive Analytics and Visualization on One Billion Tweets, Jianfeng Jia, Chen Li, Xi Zhang, Chen Li, Michael J. Carey, Simon Su, ACM SIGSPATIAL 2016 (Demo Paper)
Cloudberry Architecture
Drum: Adaptive Framework for Query Slicing

Drum: A Rhythmic Approach to Interactive Analytics on Large Data, Jianfeng Jia, Chen Li, Michael J. Carey, IEEE Big Data 2017
Tutorial Steps

● Requirements
  ○ Shell terminal
  ○ Web browser

● Google “UCI Cloudberry”
  ○ “Resources” -> “BOSS 19 Tutorial”
Under-the-hood details
Drum: Adaptive Framework for Query Slicing
Schedule cost

- Total running time
- Smoothness of result delivery

\[ \text{Cost}(S) = \text{Cost}_t(S) + \alpha \sum D_i. \]
Linear regression with uncertainty

- \( f(r) = 0.08r + 0.9 \)

![Graph showing linear regression with uncertainty](image-url)
Tradeoff of Running Time and Penalty

[Diagram showing the tradeoff between running time and penalty, with different scenarios labeled (a) to (d).]
Choosing $r_i$ to maximize the expected score