Too Big to Eat: Boosting Analytics Data Ingestion from Object Stores with Scoop

Yosef Moatti[1], Eran Rom[2], Raúl Gracia-tinedo[3], Dalit Naor[1], Doron Chen[1], Josep Sampé[3], Marc Sánchez-Artigas[3], Pedro García-López[3], Filip Gluszak[4], Eric Deschodt[4], Francesco Pace[5], Daniele Venzano[5], Pietro Michiardi[5]
IBM Research[1], OpenStack Storlets Project[2], Universitat Rovira i Virgili[3], GridPocket[4], Eurecom[5]

Problem

- Large Volumes of Unstructured Data Objects
- Disaggregated Compute and Storage Clusters
- Costly Data Ingestion When Running SQL Queries

Goals

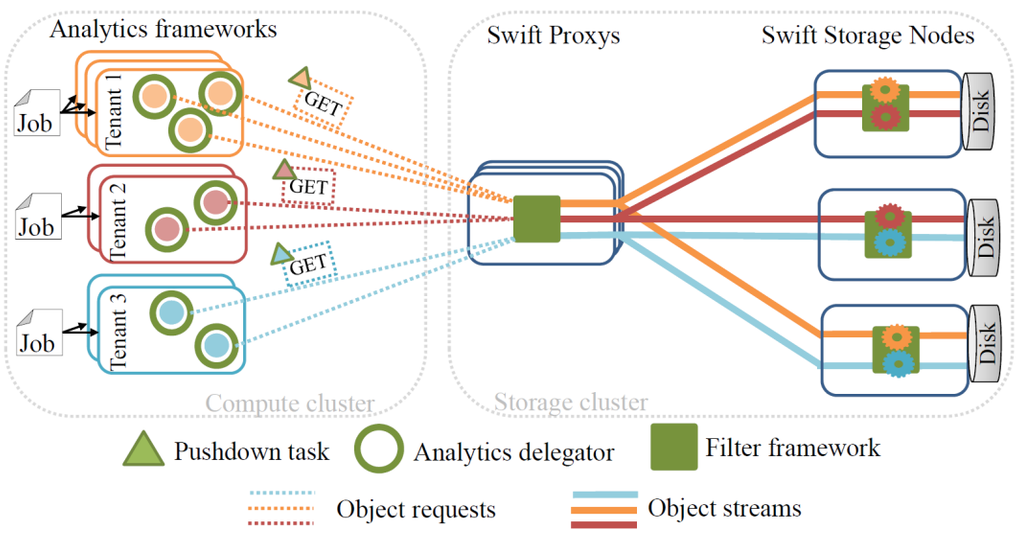
- Analytics Jobs Delegate SQL Tasks to The Storage
- The Storage Cluster Runs These Tasks To Cut Ingestion
- Framework To Orchestrate Cooperation between Analytics Jobs and Storage

Results

- Open-Source Prototype Implementation
- GridPocket Use-Case: Energy Data Management
- Evaluation on a 63-Machine Cluster with Zoe – Analytics on demand

Platform & Experiments

Conceptual Design of Scoop



Implementation/Deployment

Spark SQL

New RDD & Data Source to Delegate SQL Predicates

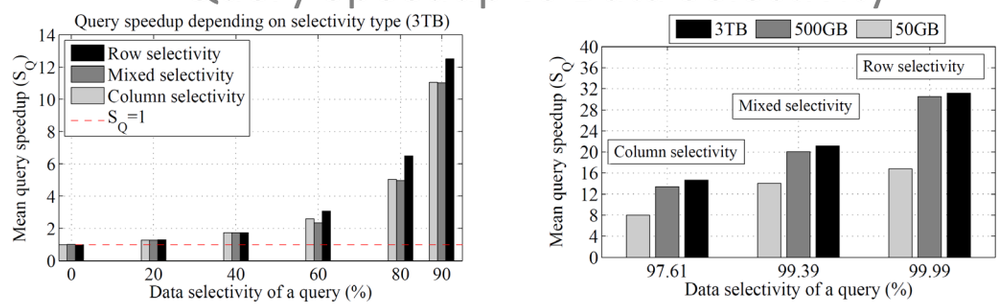
openstack. [Swift + Storlets]

New Storlet to handle Selections & Projections

25 Spark Nodes — 10 Gbps — 35 Swift Nodes

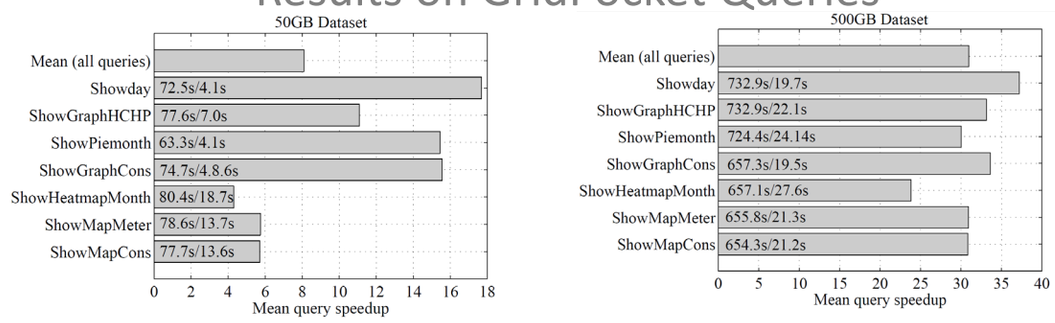
OSIC OPENSTACK INNOVATION CENTER

Query Speedup vs Data Selectivity



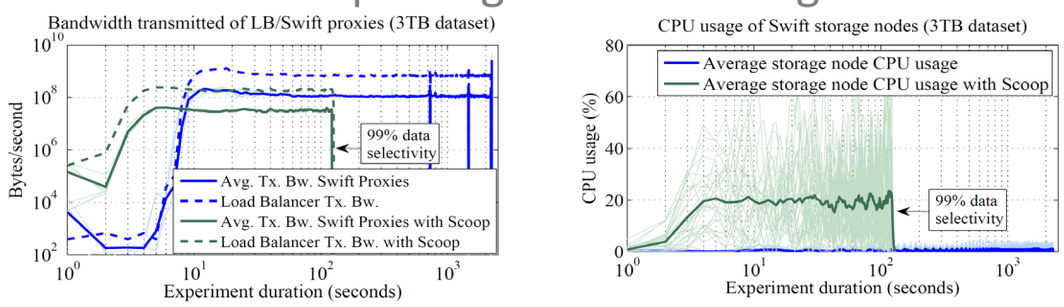
High Query Speedup: As Query Data Selectivity Grows, Job Speedup Exhibits a Non-linear Improvement

Results on GridPocket Queries



Solving GridPocket Problems: Typical GridPocket Data Intensive Queries Show Speedups Up To x32

Inspecting Resource Usage



Nice trade-off: We Avoid Consuming 99% Transfer BW by Using 20% of (Mostly Idle) Storage Compute Power

IOSTACK
SOFTWARE-DEFINED STORAGE FOR BIG DATA

Find US!

http://iostack.eu

@iostackproject

https://github.com/iostackproject

GRIDPOCKET PERSONAL SMARTGRID SOLUTIONS

UNIVERSITAT ROVIRA I VIRGILI

IBM

EURECOM Sophia Antipolis

openstack.

OSIC OPENSTACK INNOVATION CENTER

THIS PROJECT HAS RECEIVED FUNDING FROM THE EUROPEAN UNION's HORIZON 2020 RESEARCH AND INNOVATION PROGRAMME UNDER GRANT AGREEMENT NUMBER 644182