



HAL
open science

Scaling Smart Appliances for Spatial Data Synthesis

Luis Pineda-Morales, Balaji Subramaniam, Kate Keahey, Gabriel Antoniu, Alexandru Costan, Shaowen Wang, Anand Padmanabhan, Aiman Soliman

► **To cite this version:**

Luis Pineda-Morales, Balaji Subramaniam, Kate Keahey, Gabriel Antoniu, Alexandru Costan, et al.. Scaling Smart Appliances for Spatial Data Synthesis. SC15 - ACM/IEEE International Conference in Supercomputing, Nov 2015, Austin, United States. , 2015. hal-01241718

HAL Id: hal-01241718

<https://inria.hal.science/hal-01241718>

Submitted on 10 Dec 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



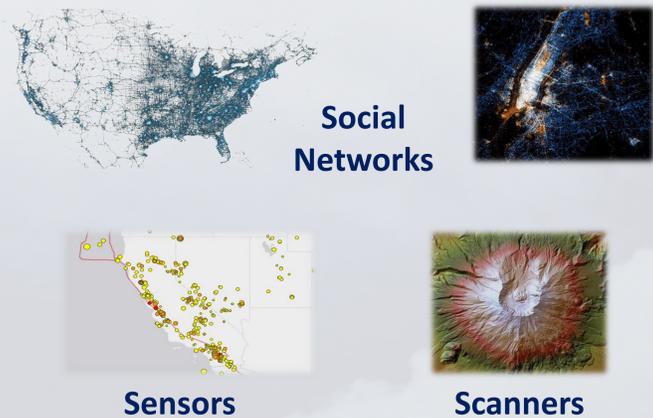
Scaling Smart Appliances for Spatial Data Synthesis

Luis Pineda-Morales^{1,2}, Balaji Subramaniam³, Kate Keahey³, Gabriel Antoniu², Alexandru Costan^{2,4}
 Shaowen Wang^{5,6,7}, Anand Padmanabhan^{5,6,7}, and Aiman Soliman^{5,6}

Problem

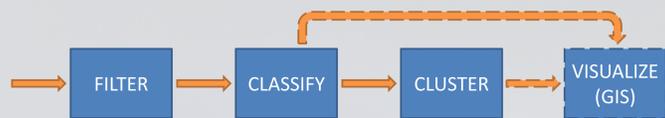
- *Spatial data streams* have grown considerably in size and complexity
- Volatility and growing number of user requests through CyberGIS Gateway [1]

Dynamic Data Streams



CyberGIS

- Software integration for sustained geospatial innovation
- Achieve *scalability* for large geospatial problems and number of users
- *Use Case*: Spatial index for estimating home and work relocation, and unemployment rates using geo-located twitter data



Challenges

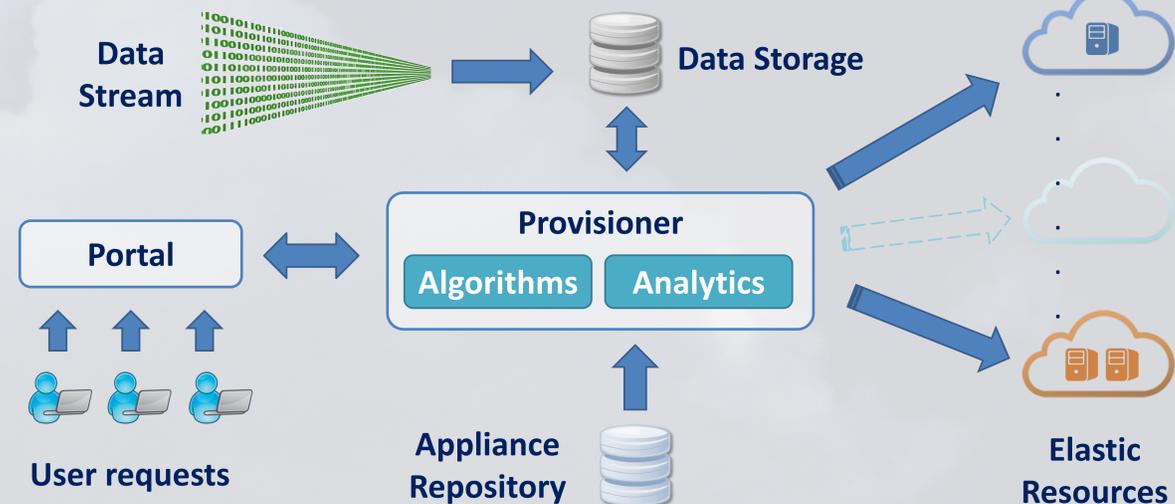
- How to deal with data and user requests volatility?
- How can we provision appliances quickly and efficiently?

Approach

- Scale the computer resources in order to meet volatility
- *Smart Appliances*



Architecture



Experimental Setup

- 2 bare-metal Chameleon nodes: 24 cores (48 VCPUs), 256 GB disk, 128 GB RAM
- Hadoop 2.7 (YARN): 128 MB default HDFS block size, up to 192 containers
- Up to 20 days of geo-located tweets (~1.9GB/day)

References

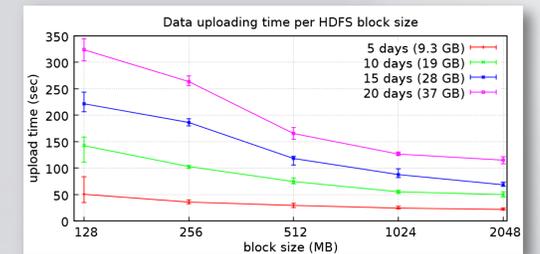
- [1] CyberGIS Software Integration for Sustained Geospatial Innovation <http://cybergis.cigi.uiuc.edu/>
- [2] Infrastructure Outsourcing in Multi-Cloud Environment, Keahey, K., et al. Workshop on Cloud Services, Federation, and the 8th Open Cirrus Summit. 2012

Acknowledgements

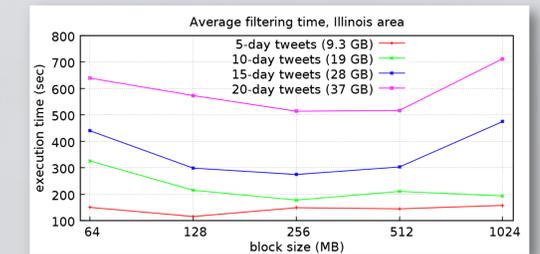
- Luis Pineda-Morales' internship at Argonne National Laboratory was funded by the Data@Exascale Inria-ANL Associate Team and by the NextGN Inria-ANL PUF project
- This material is based in part upon work supported by the U.S. National Science Foundation under grant numbers: 1047916, 1429699, and 1443080. Insightful comments and feedbacks received from the following members of the CyberGIS Center: Junjun Yin and Kiumars Soltani are greatly appreciated

Developing a Statistical Model

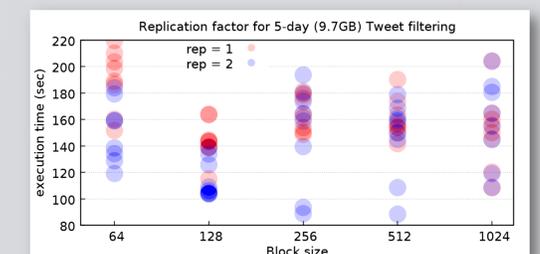
- **Data Upload**
- 70%+ improvement with larger block size, compared to the default 128 MB



- **Execution Time**
- "Sweet spot" more evident in larger datasets



- **Data Replication**
- Large dataset: disk saturation



Future Work

- Scale out to months of geo-located tweets
- Test in other [virtualized] infrastructures
- Define models for elastic appliances
- Support for new job releases of the use case involving Phantom elastic provisioner [2]

¹ Microsoft Research – Inria Joint Centre, FR
² KerData Project-Team, Inria, FR
³ Argonne National Laboratory, US
⁴ IRISA / INSA Rennes, FR

⁵ CyberGIS Center for Advanced Digital and Spatial Studies, UIUC, US
⁶ National Center for Supercomputing Applications, UIUC, US
⁷ Department of Geography and Geographic Information Science, UIUC, US

