A Toolkit for Reproducible Big Data Analytics in the Cloud

2022 ESIP January Meeting

Xin Wang (xinwang11@umbc.edu), Jianwu Wang (jianwu@umbc.edu)
University of Maryland, Baltimore County

Our Reproducible and Portable Big Data Analytics in the Cloud (RPAC) Toolkit

- By leveraging serverless, containerization and adapter design pattern techniques, our RPAC toolkit can achieve reproducibility, portability and scalability for big data analytics in the cloud

Demo Outline

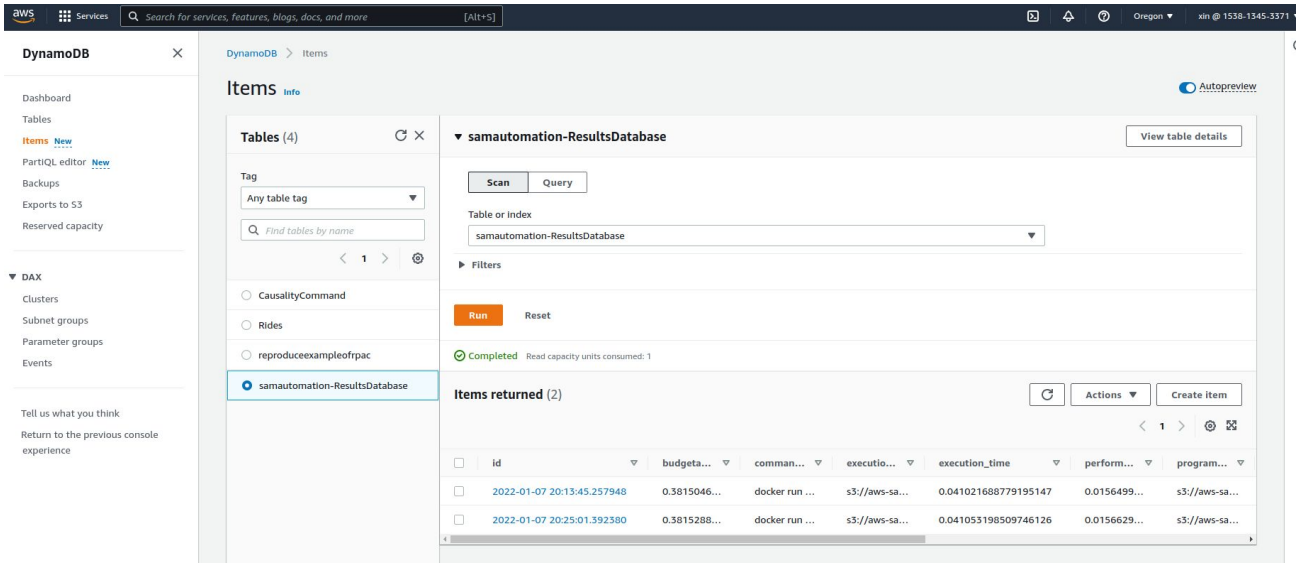
- One-click execution for Cloud Retrieval application
- Showing execution results from database and Cloud storage
- Reproduce from execution history

One-click Execution for Cloud Retrieval Application

1. Check out our code from GitHub repository: https://github.com/big-data-lab-umbc/Reproducible_and_portable_app_in_cloud
2. Update *ConfigTemplate/personal.ini* file based on your cloud credentials
3. Check RPAC usage
    - *python3 main.py -h*
4. One-click execution for Cloud Retrieval application
    - *python3 main.py --one_click*

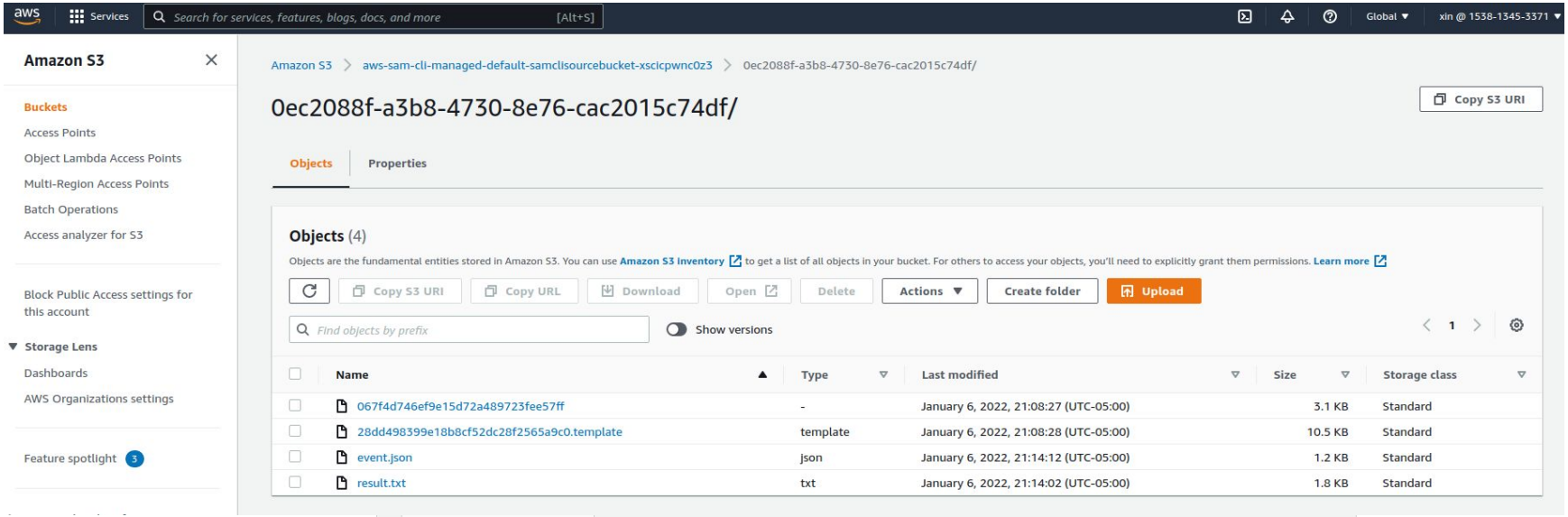
Showing Execution Results from Database

- Meta data of the execution result will be shown in DynamoDB on AWS:



Showing Execution Results from Cloud Storage

- Detailed execution result will be shown on AWS S3



Reproduce from Execution History

1. Download the execution history folder from S3 to local folder *ExecutionHistory*: https://github.com/big-data-lab-umbc/Reproducible_and_portable_app_in_cloud
2. Update ConfigTemplate/personal.ini file based on your cloud credentials
3. Command to reproduce from execution history
    - *python3 main.py --execution_history 0ec2088f-a3b8-4730-8e76-cac2015c74df --one_click*

Detailed Information

- Paper and GitHub repository of our toolkit:
  https://bdal.umbc.edu/tools/#reproducible-data-analytics

Thank you very much! :)