



A Toolkit for Reproducible Big Data Analytics in the Cloud

2022 ESIP January Meeting

[Xin Wang \(xinwang11@umbc.edu\)](mailto:xinwang11@umbc.edu), [Jianwu Wang \(jianwu@umbc.edu\)](mailto: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](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

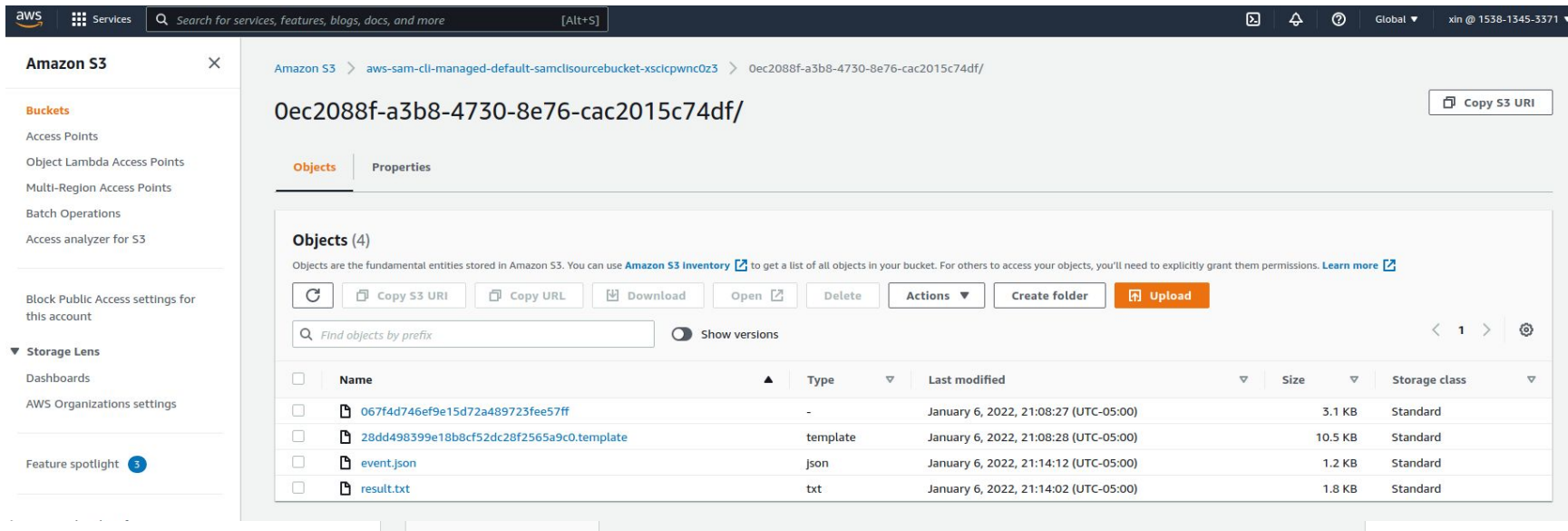
- Meta data of the execution result will be shown in DynamoDB on AWS:

The screenshot shows the AWS Management Console interface for the 'samautomation-ResultsDatabase' table in DynamoDB. The left sidebar contains navigation links for Dashboard, Tables, Items, PartiQL editor, Backups, Exports to S3, Reserved capacity, DAX, Clusters, Subnet groups, Parameter groups, and Events. The main content area displays the 'Items' view for the selected table. It includes a 'Scan' button, a 'Query' button, and a 'View table details' link. Below these, there is a 'Table or index' dropdown menu set to 'samautomation-ResultsDatabase'. A 'Filters' section is visible. The 'Run' button is highlighted, and a status message indicates 'Completed' with 'Read capacity units consumed: 1'. The 'Items returned (2)' section shows a table with two rows of data.

id	budgeta...	comman...	executio...	execution_time	perform...	program...
2022-01-07 20:13:45.257948	0.3815046...	docker run ...	s3://aws-sa...	0.041021688779195147	0.0156499...	s3://aws-sa...
2022-01-07 20:25:01.392300	0.3815288...	docker run ...	s3://aws-sa...	0.041053198509746126	0.0156629...	s3://aws-sa...

Showing Execution Results from Cloud Storage

- Detailed execution result will be shown on AWS S3



The screenshot shows the AWS S3 console interface. The top navigation bar includes the AWS logo, 'Services' link, a search bar, and user information. The left sidebar shows the 'Amazon S3' section with various options like Buckets, Access Points, and Storage Lens. The main content area displays the details of a bucket named 'aws-sam-cli-managed-default-samclisourcebucket-xscicpwn0z3' and a specific folder '0ec2088f-a3b8-4730-8e76-cac2015c74df/'. The folder is currently selected, showing a list of objects. The objects are:

Name	Type	Last modified	Size	Storage class
067f4d746ef9e15d72a489723fee57ff	-	January 6, 2022, 21:08:27 (UTC-05:00)	3.1 KB	Standard
28dd498399e18b8cf52dc28f2565a9c0.template	template	January 6, 2022, 21:08:28 (UTC-05:00)	10.5 KB	Standard
event.json	json	January 6, 2022, 21:14:12 (UTC-05:00)	1.2 KB	Standard
result.txt	txt	January 6, 2022, 21:14:02 (UTC-05:00)	1.8 KB	Standard



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! :)