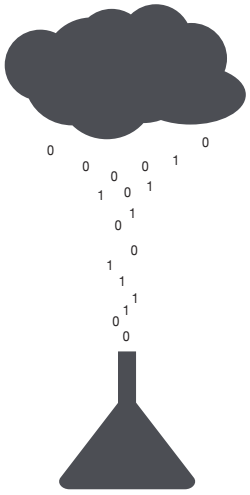
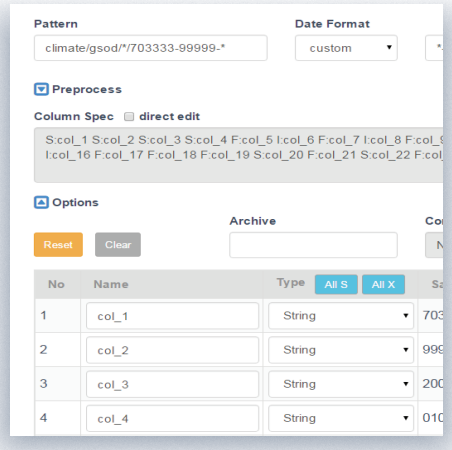


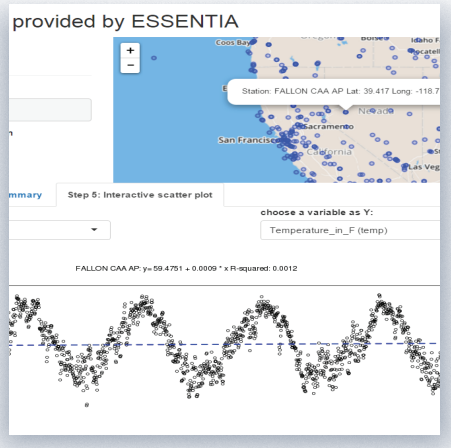
Essentia Labs : Analyzing Global Weather Measurements from the GSOD for the period 1929 through 2009

The Essentia data science team wanted to demonstrate how easy it is for anyone with a little bit of scripting knowledge can immediately gain value from the large pools of open data sets using Essentia in the cloud. They decided to use the public dataset **Daily Global Weather Measurements, 1929-2009 (NCDC, GSOD)** available for free on Amazon Web Services. At a total data size of 20 gb, they estimated only a single Essentia instance would be necessary to perform all the necessary processing and analysis.

Connecting to the Data

First the team launched an m3 medium EC2 instance on AWS with the Essentia AMI installed. The data was already publicly available on AWS as an EBS volume, so they just had to copy the snapshot to an EBS mounted to our EC2 instance, then transfer the files to an S3 bucket. From the Essentia UI, they added the S3 bucket as a data repository, and that's it.





Exploring the data

Using the data viewer features, they browsed the directory and files to see how many files they were dealing with and overall data size.
Then they made some exploratory data categories to get an idea about the structure and types of data.
This was done without performing any ETL, or having to move data out of S3.

Preparing data

For their demo, they wanted to focus on US based weather stations.
Using the command line tools, they created a script that filtered out all non-us weather stations based on country code and then removed any stations that did not contain any data.
Once the script is executed, all the target data is then loaded into Essentia's in-memory DB.

Analyzing the data in R

Data can now be streamed into an R dataframe for statistical analysis and then visualized through Shiny apps which has a variety of different visualization features including maps, charts, and plots.

This is just a simple example of what can be accomplished using Essentia.
In addition, the data science team could have mashed up with other data sets or implemented some predictive or machine learning algorithms for future forecasting and modeling.

Find out more at *https://github.com/auriq/ShinyApps* for detailed instructions and script samples.

Essentia + Open Data

Thousands of open data sets have been made available over the past few years through various public and private organizations, allowing anyone to potentially make use of the data by analyzing and presenting their results in unique and meaningful ways. A common problem when dealing with public or open data is that the formats vary from source to source, it is often stored as compressed files, and the overall volume of data can be incredibly large making it difficult to even begin analyzing and producing results.

The Essentia platform is the perfect solution for any data scientist, researcher, or analyst to be able to immediately gain value from the multitude of open data feeds that are publicly available. It reduces the technical hurdles that anyone working with open data experiences including data ingestion, exploration, preparation, visualization and management.

Data ingestion

Essentia in the cloud was designed to work with distributed cloud data stores like Amazon S3 or MS Azure Blob. Simply dump the data without any pre-conditioning into these repositories through any means you desire, including FTP, Flume, Sqoop, API, S3 Copy, etc. In many cases, there are already S3 buckets freely available with the various public data sets already loaded. If the files are in compressed format, there is no need to decompress or convert to other formats, since Essentia was designed to handle multiple file compression schemes.

Data exploration

Essentia allows you to browse your repositories, scan and sample files, create data categories, run SQL like queries directly on the raw data without having to move the data outside of repositories. This streamlines initial data exploration jobs and eliminates data movement costs.

Data Visualization

Essentia's integration with R makes it easy to turn your data into amazing interactive charts by using Shiny apps to visualize your data. You can also directly output your data to load into Tableau, Qlikview or any other popular visualization tool.

Data preparation

Use Essentia commands to perform data preparation at scale including data cleansing, transformation, and blending. Similar to bash script, Essentia scripting language is easy to learn and use. Essentia supports many high level languages, such as Python, to programmatically execute combinations of commands. Integration with R allows you to easily prepare and load your data into an R data frame for complex statistical analysis. Or just load your prepared data into Amazon Redshift or other relational database.

Data Management

Essentia makes it easy to manage open data. Regardless if your data is all in one repository, or in multiple repositories across platforms (AWS or Azure), you can easily connect to all of them with Essentia. Organize your data into virtual categories that allow you to work with data without having to waste time trying to change or modify the original raw data. No need to move data into or out of their existing locations.

Get Essentia on AWS

1. Search for "Essentia HVM" from AWS Marketplace
2. Select EC2 instance type to install Essentia machine image (free to use on t2.micro)
3. Launch and login to UI and begin exploring and analyzing your data instantly

For more information go to *http://www.auriq.com/documentation/source/install*