Apache Beam Tutorial – Learn Beam API for Big Data Ecosystem

Apache Beam

Apache Beam is an open source from Apache Software Foundation. It is an unified programming model to define and execute data processing pipelines. The pipelines include ETL, batch and stream processing. Apache Beam has published its first stable release, 2.0.0, on 17th March, 2017. There is active development around Apache Beam from Google and Open Community from Apache.

MapReduce has triggered the evolution of Big Data Ecosystem that we are seeing today. There are many frameworks like Hadoop, Spark, Flink, Google Cloud Dataflow, etc, that came into existence. But there has been no unified API that binds all these frameworks and data sources, and provide an abstraction to the application logic from big data ecosystem. The good days have come, Apache Beam framework provides abstraction between your application logic and big data ecosystem.

Hence, there is no need to bother about the following aspects when you are writing your data processing or analytic application :

- **Data Source** – Data source can be batches, micro-batches or streaming data
- **SDK** – You may choose your SDK (Java, Python) that you are comfortable with, to program application logic as a Beam Pipeline
- **Runner** – Once writing of application logic as a Beam Pipeline is done, you may choose one of the available runners (Apache Spark, Apache Flink, Google Cloud Dataflow, Apache Apex, etc.) to run your application based on the nature of your inputs and analytic results.

This is how, Beam lets you write your application logic once, and not mix and scramble the code with input specific parameters or runner specific parameters.

Apache Big Data Tutorials

- Learn Apache Hadoop
- Learn Apache Spark
- Learn Apache Flink
- Learn Apache Beam
- Learn Apache MXNet

Apache Database Projects

- Apache CouchDB Tutorial

Apache Other Projects

- Apache Tomcat Tutorial
- Apache PDFBox Tutorial
- Learn Apache OpenNLP