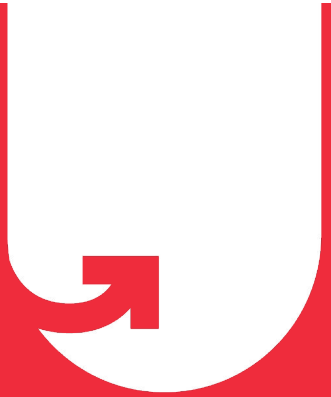
upGrad

BATCH PROCESSING WITH APACHE FLINK

About

upGrad



upGrad

Course: Data Engineering - II

Lecture On: Apache Flink

Instructor: Mayukh Chakraborty

upGrad

INTRODUCTION TO DATASET API

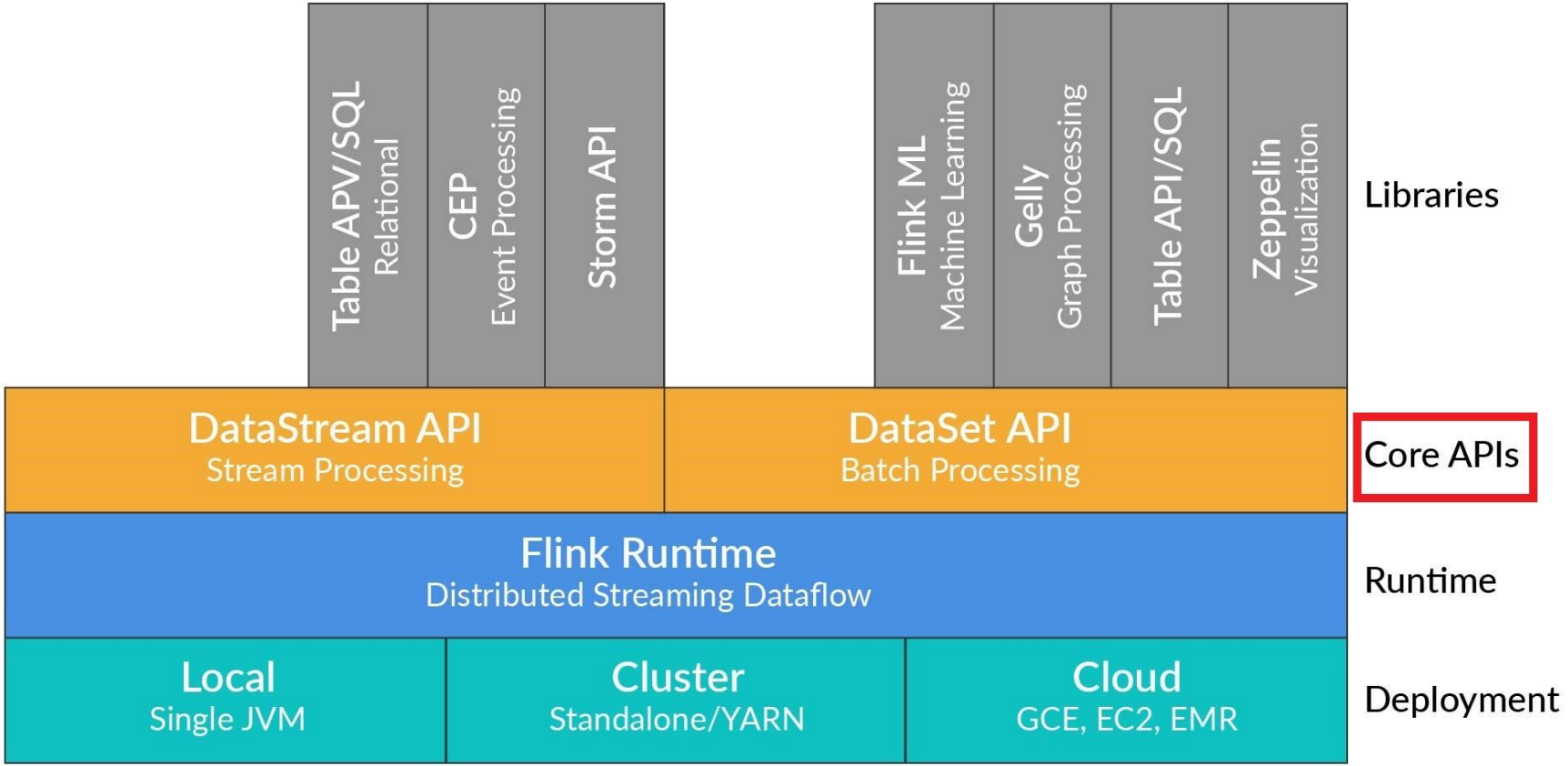
IN THIS SEGMENT

Understand the concept of a Dataset and Tuple.

Understand the anatomy of a Flink program.

Try a simple Flink program in Java.

COMPONENT STACK



DataSet

01	It is a finite, immutable collection of Data objects.
02	It can contain duplicates.
03	Data is read from the source into DataSet and in every transformation step, a new Dataset gets created.

```
DataSet<String> words = …
```

TUPLE

01 It is a finite collection of data attributes.

02 Holds different attributes of a single data object.

03 Apache Flink has defined Tuple0, Tuple1, Tuple2 … upto Tuple25

```
DataSet<Tuple2<String, Integer>> wordCounts = …
```

ANATOMY OF A FLINK PROGRAM

1. Obtain an execution environment.
2. Initially load data from the data source.
3. Specify transformations on this data.
4. Specify the data sink.
5. Trigger the program execution.

Validate Email addresses using Apache Flink DataSet

upGrad

TRANSFORMATIONS

IN THIS SEGMENT

Learn about various types of transformations available with Dataset API.

How transformations can be used in Flink programs?

MAP

Takes element in one format and transform into another format.

```
data.map(new MapFunction<String, Integer>() {
      public Integer map(String value) {
          return Integer.parseInt(value); } });
```

FlatMap

Takes one element and produces zero, one, or more elements.

```
data.flatMap(new FlatMapFunction<String, String>() {
      public void flatMap(String value, Collector<String> out) {
          String[] tokens = value.toLowerCase().split("\\W+");
          for (String token : tokens) {
              out.collect(new Tuple2<String, Integer>(token, 1));}}
  });
```

FILTER

Filter out values, for which the expression predicate doesn't meet.

```
data.filter(new FilterFunction<String>() {
      public boolean filter(String email) {
          return isValidEmail(email); }
  });
```

UNION

Produces the union of two data sets.

```
DataSet<String> result = data1.union(data2);
```

JOIN

Joins two data sets in a manner similar to SQL.

```
result = input1.join(input2)
     .where(0)     // key of the first input(tuple field 0)
     .equalTo(1);  // key of the second input(tuple field 1)
```

upGrad

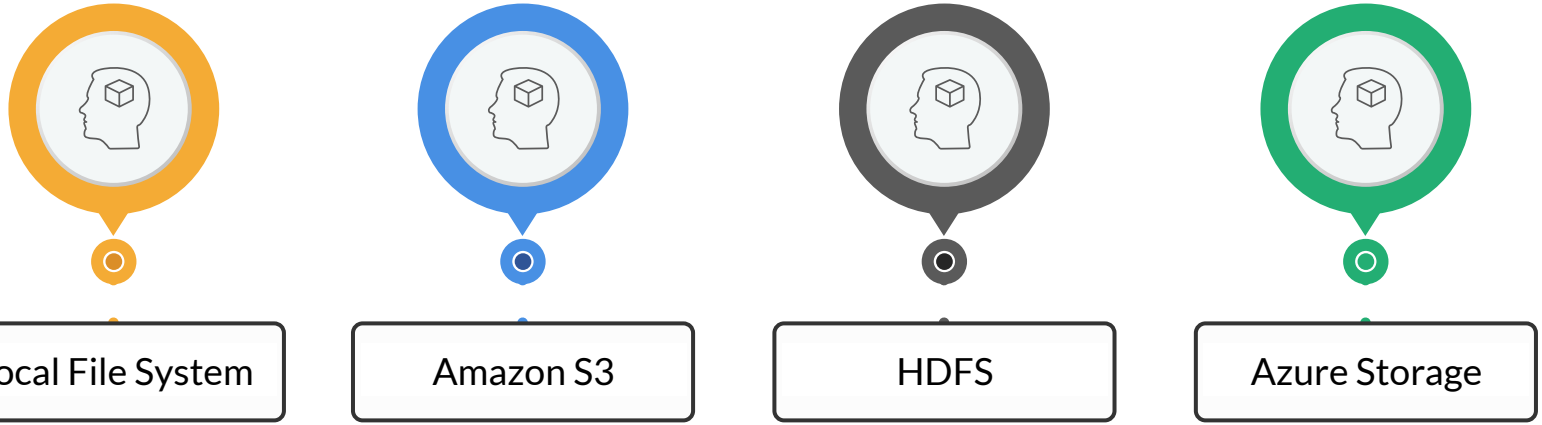
CONNECTORS

IN THIS SEGMENT

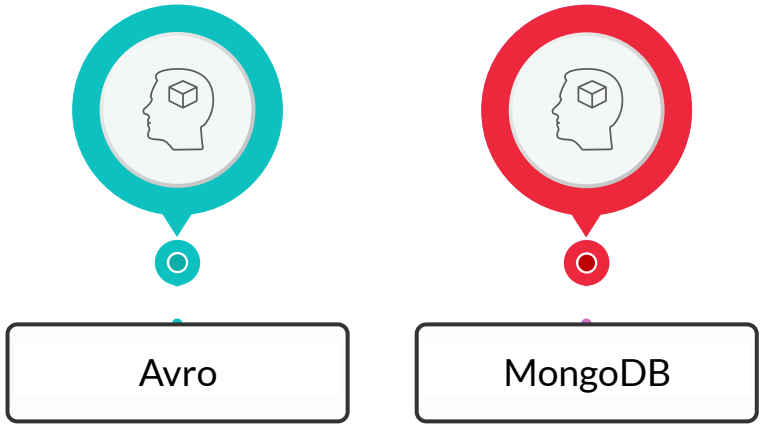
Know about various types of connectors which can be used with Dataset API.

How connectors can be used in Flink programs?

TYPES OF CONNECTORS



TYPES OF CONNECTORS



upGrad

Exercise