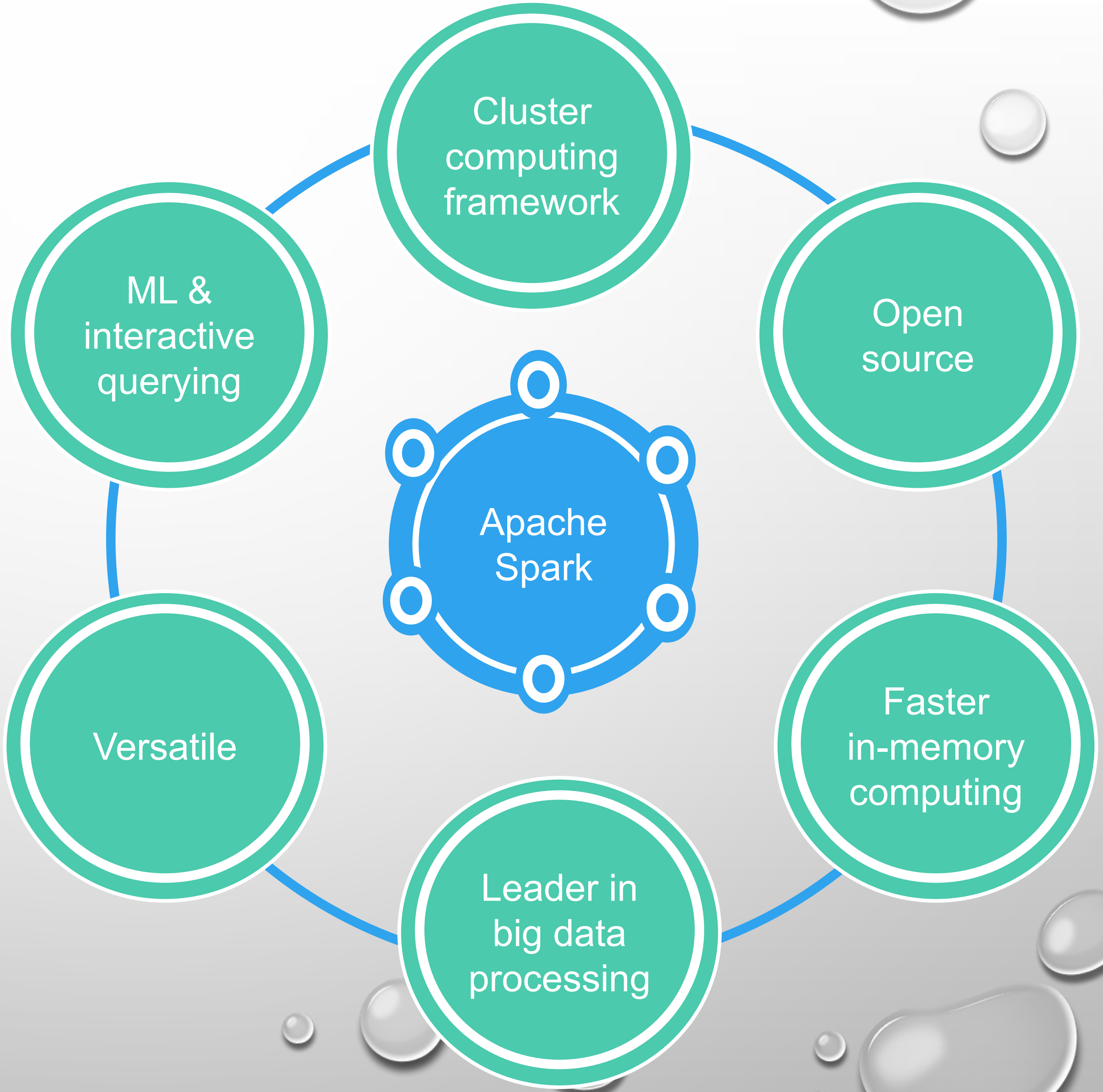
APACHE SPARK

CONTENTS

- APACHE SPARK
- SPARK ARCHITECTURE
- RDD
- DATAFRAME API
- SPARKSQL
- EDA WITH PYSPARK
- PREDICTIVE ANALYSIS WITH SPARK MLIB

Apache Spark is a distributed processing system used to perform big data and machine learning tasks on large datasets.



Apache Spark

Cluster computing framework

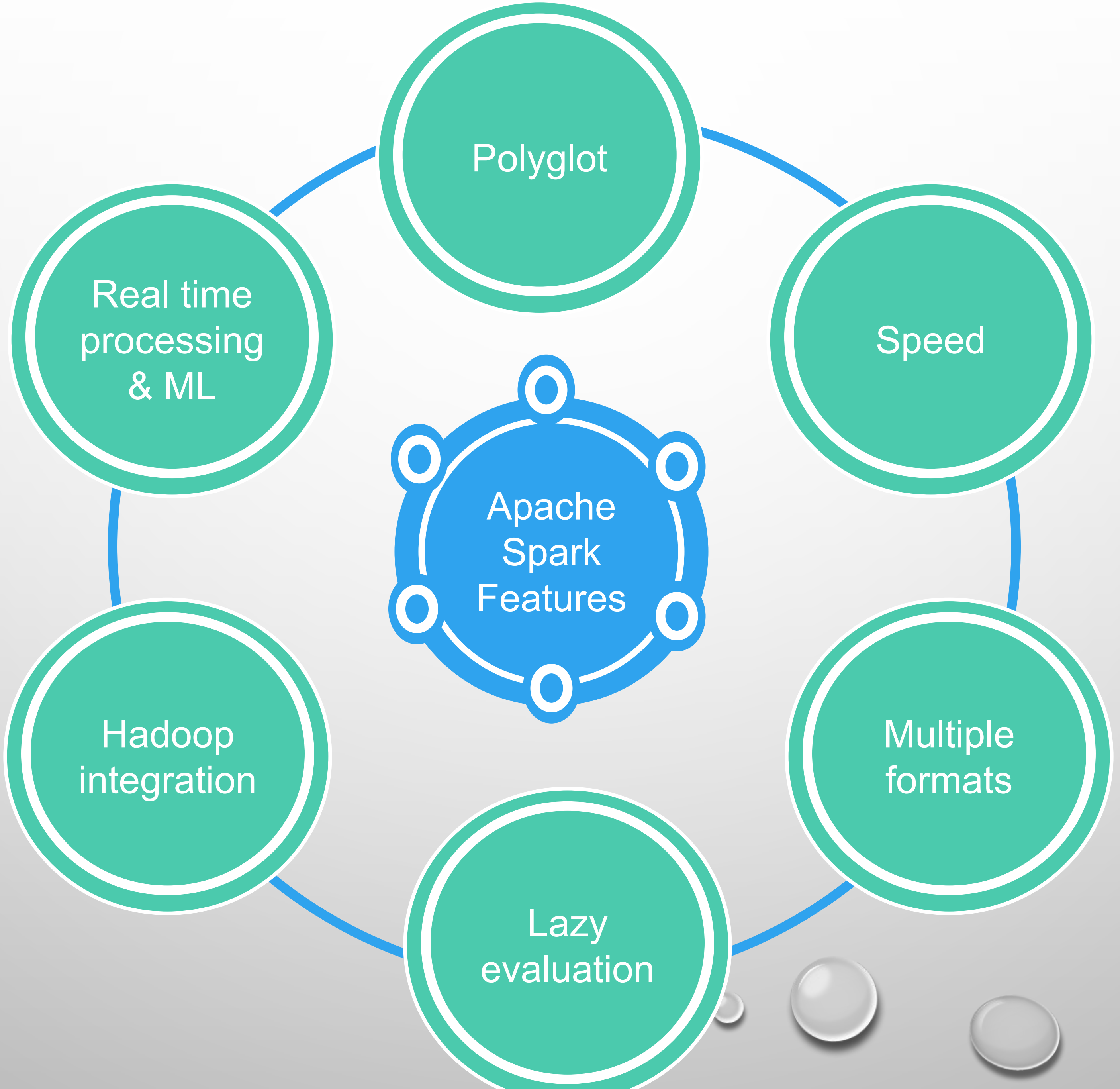
Open source

Faster in-memory computing

Leader in big data processing

Versatile

ML & interactive querying



Apache Spark Features

Polyglot

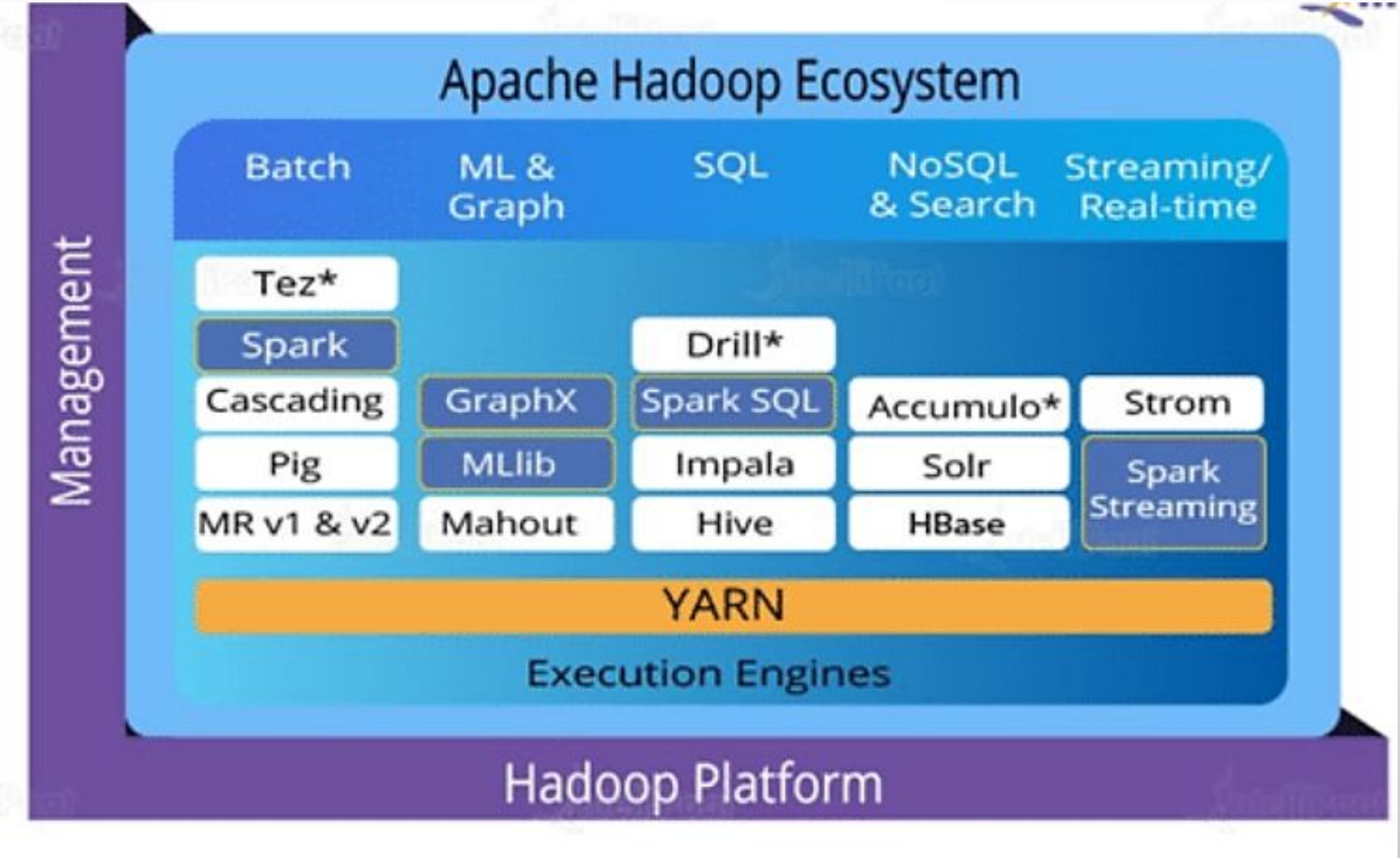
Speed

Multiple formats

Lazy evaluation

Hadoop integration

Real time processing & ML



Apache Hadoop Ecosystem

Batch	ML & Graph	SQL	NoSQL & Search	Streaming/ Real-time
Tez*				
Spark		Drill*		
Cascading	GraphX	Spark SQL	Accumulo*	Strom
Pig	MLlib	Impala	Solr	Spark Streaming
MR v1 & v2	Mahout	Hive	HBase	

YARN

Execution Engines

Management

Hadoop Platform

PySpark Python interface to spark

Distributed processing environment

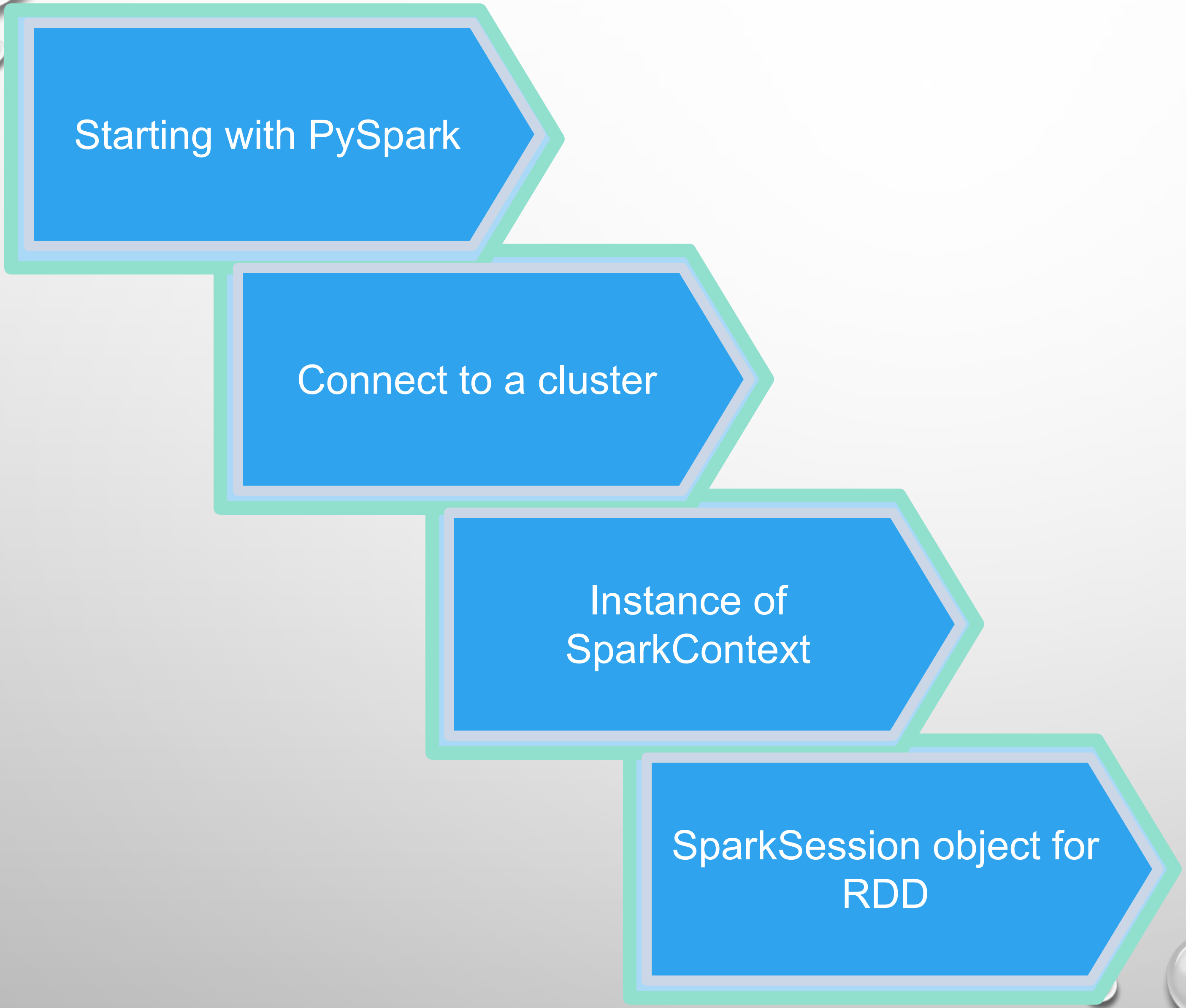
Implement machine learning workflows

Process big data

Code to collect data from a source that is continuously updated

Being around longer, better community support

Fault tolerance, can run without drive/ SSD



Starting with PySpark

Connect to a cluster

Instance of SparkContext

SparkSession object for RDD