Data Engineering Batch 1 **Day 11:03/02/2024**

Name: Pradip Bochare

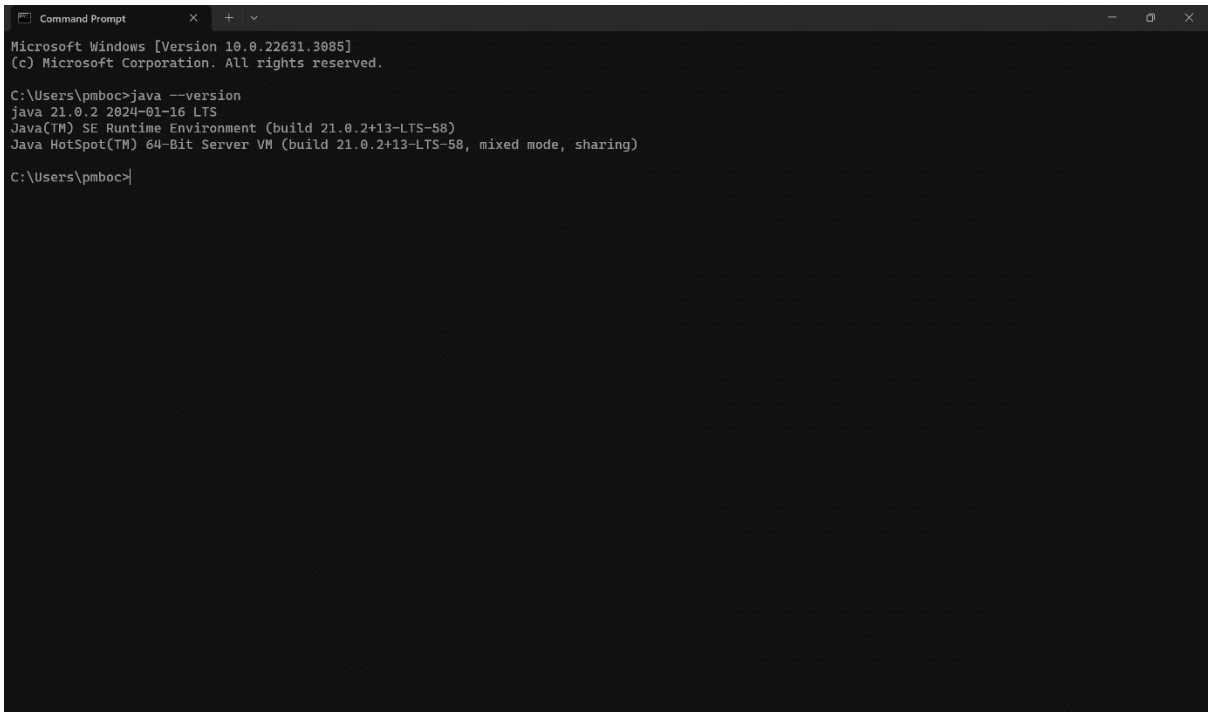
Installation of Apache Spark and Setup

❖ Apache PySpark

Apache Spark is an open-source distributed computing system that provides a fast and general-purpose cluster-computing framework for big data processing. PySpark is the Python API for Apache Spark, allowing you to write Spark applications using Python.

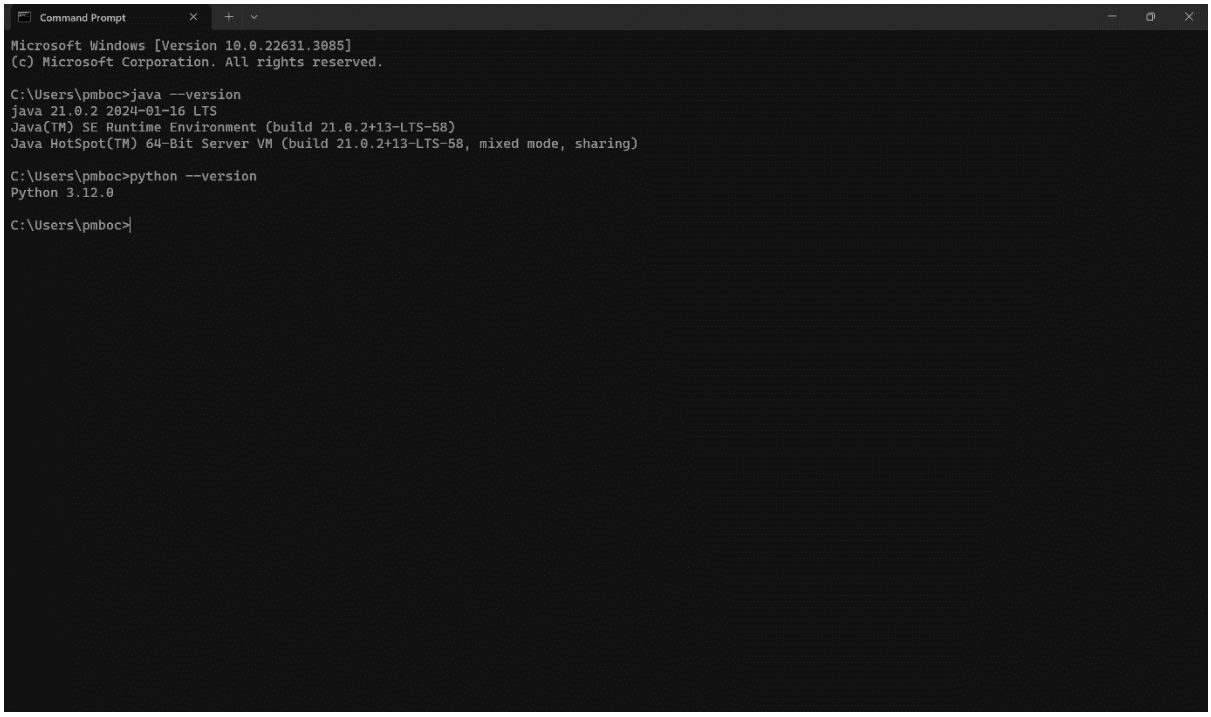
❖ Install Java

- Firstly, we have to install java jdk version which is compatible to your system. It's good to download and install latest standard version of java.
- After that we have to check in command prompt by typing java –version which displays the version of java which you downloaded.
- In means that you have successfully downloaded and installed in your system.



❖ Install Python

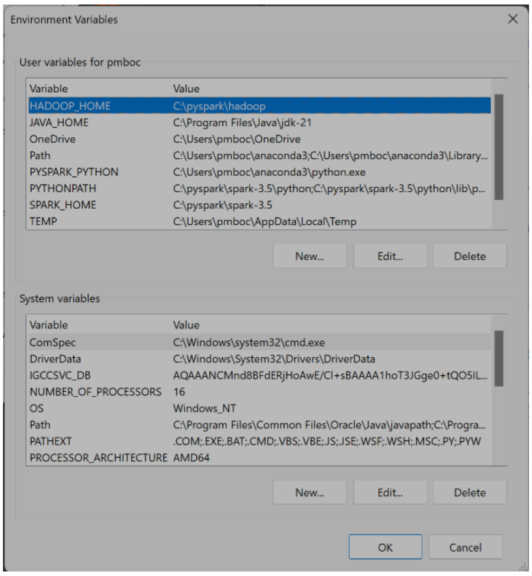
- Along with java, we have to install python environment into our system. It's good to download and install latest and standard version of python.
- After that we have to check in command prompt by typing python –version which displays the version of java which you downloaded.
- In means that you have successfully downloaded and installed in your system.

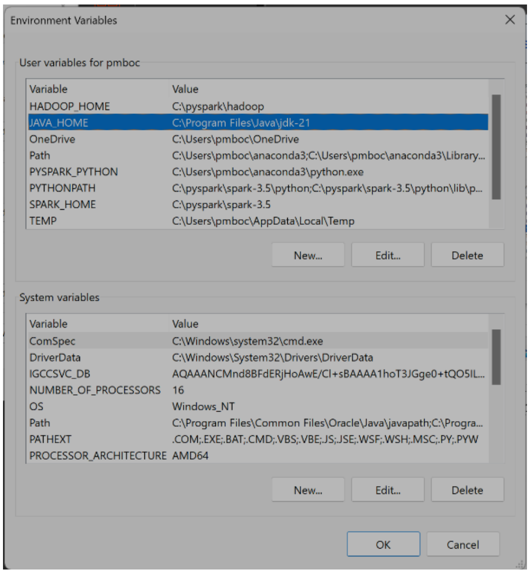


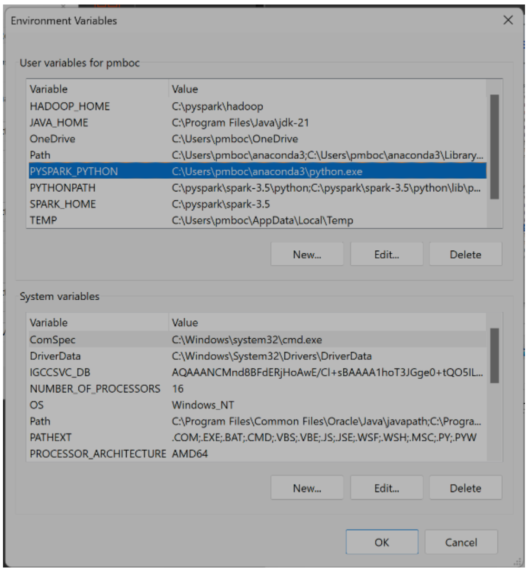
❖ Install Apache Spark

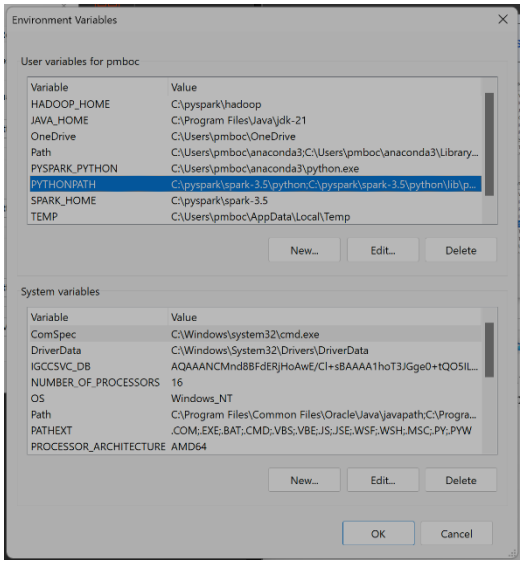
- Visit the Apache Spark download page.
- Choose the latest version of Spark and download the pre-built package for Hadoop. It will be a tarball (.tgz) file.
- Extract the downloaded tarball to a location on your machine.

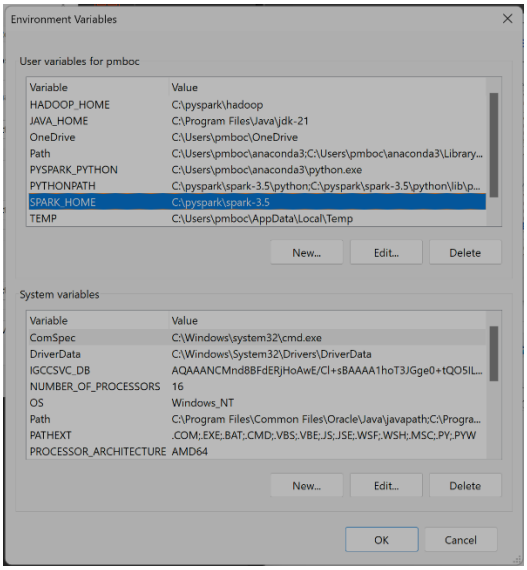
- Set Environment Variables

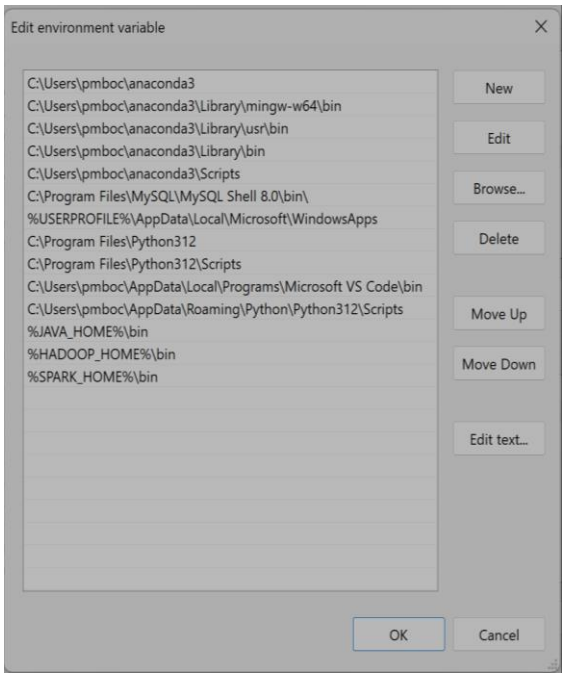




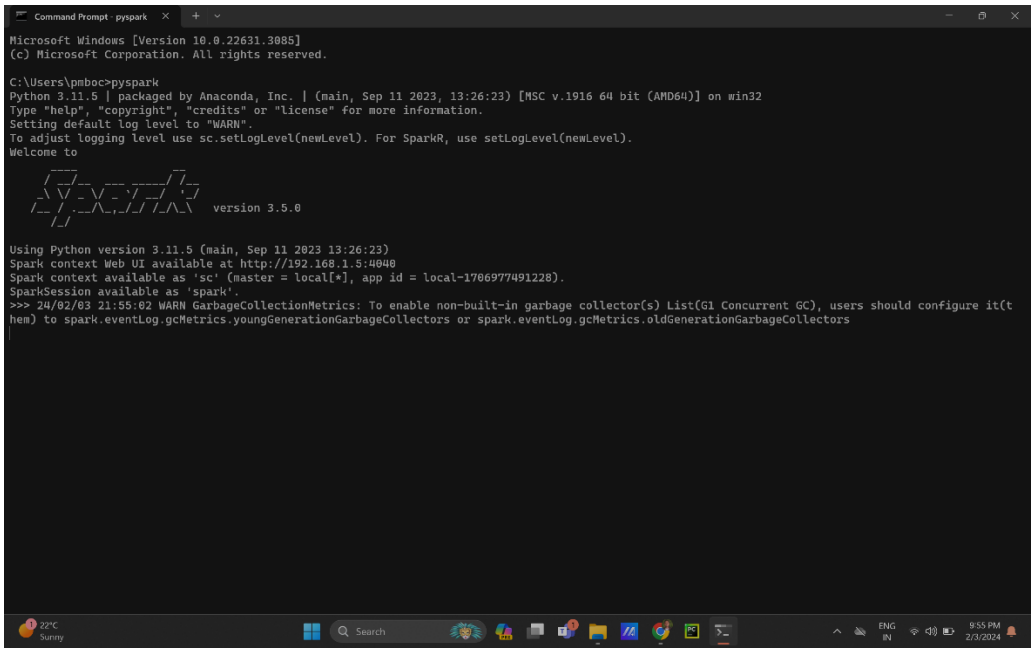



Environment Variables
User variables for pmboc
Variable Value
HADOOP_HOME C:\pyspark\hadoop
JAVA_HOME C:\Program Files\Java\jdk-21
OneDrive C:\Users\pmboc\OneDrive
Path C:\Users\pmboc\anaconda3;C:\Users\pmboc\anaconda3\Library...
PYSPARK_PYTHON C:\Users\pmboc\anaconda3\python.exe
PYTHONPATH C:\pyspark\spark-3.5\python;C:\pyspark\spark-3.5\python\lib\p...
SPARK_HOME C:\pyspark\spark-3.5
TEMP C:\Users\pmboc\AppData\Local\Temp
New... Edit... Delete
System variables
Variable Value
ComSpec C:\Windows\system32\cmd.exe
DriverData C:\Windows\System32\Drivers\DriverData
IGCCSVC_DB AQAAANCMnd8BFdERjHoAwE/Cl+sBAAAA1hoT3JGge0+tQO5IL...
NUMBER_OF_PROCESSORS 16
OS Windows_NT
Path C:\Program Files\Common Files\Oracle\Java\javapath;C:\Progra...
PATHEXT .COM;.EXE;.BAT;.CMD;.VBS;.VBE;.JS;.JSE;.WSF;.WSH;.MSC;.PY;.PYW
PROCESSOR_ARCHITECTURE AMD64
New... Edit... Delete
OK Cancel

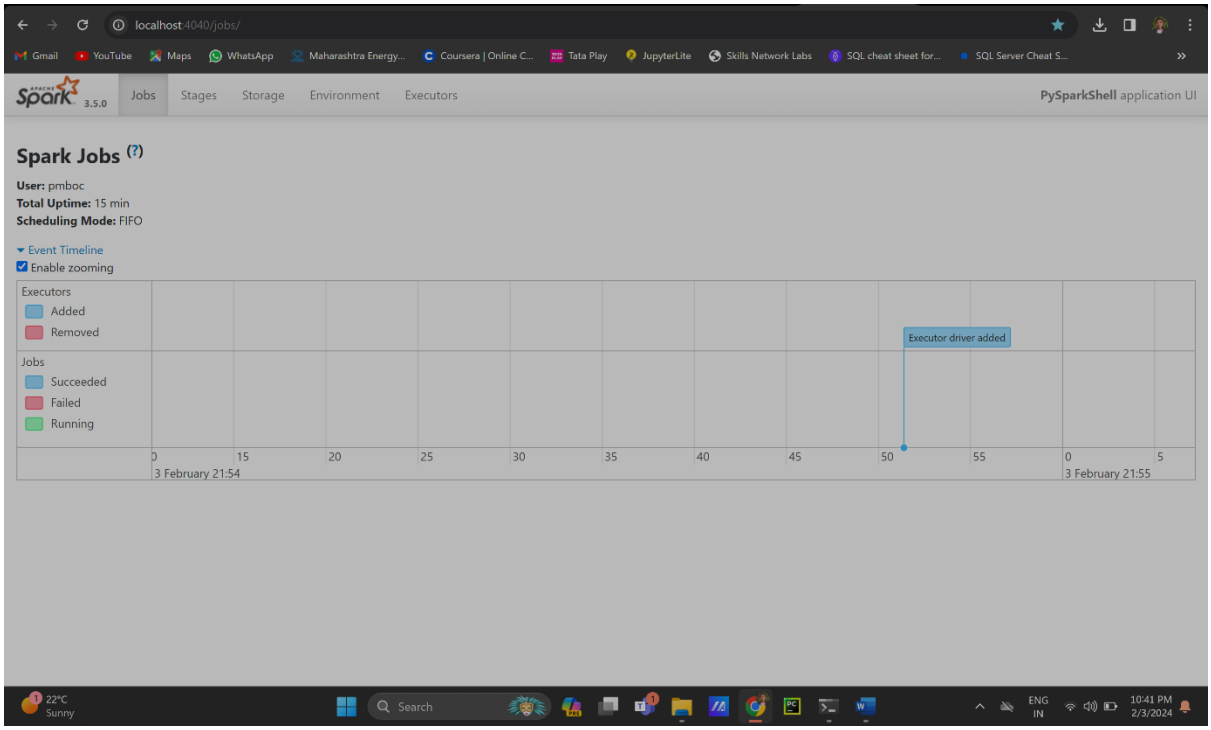

Environment Variables
User variables for pmboc
Variable Value
HADOOP_HOME C:\pyspark\hadoop
JAVA_HOME C:\Program Files\Java\jdk-21
OneDrive C:\Users\pmboc\OneDrive
Path C:\Users\pmboc\anaconda3;C:\Users\pmboc\anaconda3\Library...
PYSPARK_PYTHON C:\Users\pmboc\anaconda3\python.exe
PYTHONPATH C:\pyspark\spark-3.5\python;C:\pyspark\spark-3.5\python\lib\p...
SPARK_HOME C:\pyspark\spark-3.5
TEMP C:\Users\pmboc\AppData\Local\Temp
New... Edit... Delete
System variables
Variable Value
ComSpec C:\Windows\system32\cmd.exe
DriverData C:\Windows\System32\Drivers\DriverData
IGCCSVC_DB AQAAANCMnd8BFdERjHoAwE/Cl+sBAAAA1hoT3JGge0+tQO5IL...
NUMBER_OF_PROCESSORS 16
OS Windows_NT
Path C:\Program Files\Common Files\Oracle\Java\javapath;C:\Progra...
PATHEXT .COM;.EXE;.BAT;.CMD;.VBS;.VBE;.JS;.JSE;.WSF;.WSH;.MSC;.PY;.PYW
PROCESSOR_ARCHITECTURE AMD64
New... Edit... Delete
OK Cancel


Edit environment variable
C:\Users\pmboc\anaconda3
C:\Users\pmboc\anaconda3\Library\mingw-w64\bin
C:\Users\pmboc\anaconda3\Library\usr\bin
C:\Users\pmboc\anaconda3\Library\bin
C:\Users\pmboc\anaconda3\Scripts
C:\Program Files\MySQL\MySQL Shell 8.0\bin\
%USERPROFILE%\AppData\Local\Microsoft\WindowsApps
C:\Program Files\Python312
C:\Program Files\Python312\Scripts
C:\Users\pmboc\AppData\Local\Programs\Microsoft VS Code\bin
C:\Users\pmboc\AppData\Roaming\Python\Python312\Scripts
%JAVA_HOME%\bin
%HADOOP_HOME%\bin
%SPARK_HOME%\bin
New
Edit
Browse...
Delete
Move Up
Move Down
Edit text...
OK
Cancel

After setting up the environment variables, we need to save all of them and have to go to the command prompt and type **pyspark** as below.



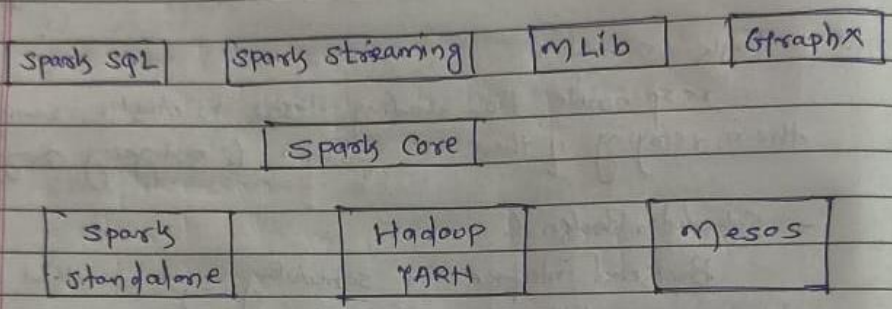
- If it shows like these, you have successfully set up the Apache Spark setup.
- If you want to see pyspark web UI, you have to type in chrome as like
- http://localhost:4040/jobs/ you will get the web page as follows which indicates your pyspark is installed as follows



Notes

* Apache Spark.
  - general purpose cluster computing system
  - provide optimized engine that supports general execution graph

• Spark Components.



* Apache Spark Core
  - All functionalities built on top of spark core
  - delivers speed by providing in-memory computation capability.

Key features
- in charge of essential I/O functionalities.
- significant in programming & observing role of spark cluster.
- Task dispatching.
- Fault recovery
- Overcomes snag of MapReduce by using in-memory computation

RDD → Resilient distributed dataset.

- spark RDD handles partitioning data across all the nodes in a cluster.

two operations performed on RDD

- Transformation
produces new RDD from existing RDD

Action.
Work with actual dataset.

* Apache spark SQL
- distributed framework for structured data processing.
- It does not depend on API / language to express the computation.
- Works to access structured & semi-structured information.

features
- cost based optimizer.
- mid query fault tolerance
- full compatibility with existing Hive data.
- Dataframes & SQL provide common way to access variety of data sources.
- provision to carry structured data inside spark programs using SQL or familiar Data Frame API

Date

* Apache Spark Streaming.
- Allows scalable, high-throughput, fault-tolerance stream processing of live data streams.

(a) Gathering
- Basic sources: file systems, socket connections
- Advanced sources: Kafka, Flume, Kinesis available through extra utility classes.

(b) Processing
- gathered data is processed using complex algorithms.

(c) Data Storage
- processed data is pushed out to file systems, databases and live dashboards.

* Apache Spark MLlib (machine Learning library)
- MLlib in spark is a scalable machine learning library that discusses both high-quality algorithm & high speed.
- clustering, regression, classification & collaborative filtering.

- Spark Context
  represent connection to a Spark cluster.

- DAG scheduler
  Computes DAG of stages for each job and submits them to TaskScheduler determines preferred locations for tasks & finds minimum schedule to run the jobs

- Task scheduler:
  responsible for sending tasks to cluster, running them, retrying if there are failures & mitigating stragglers

- Scheduler Backend.
  Backend interface for scheduling systems that allows plugging in different implementations.

- spark standalone.

RDD objects → DAG Scheduler → Task Scheduler → Worker

build operator DAG	split graph into stages of tasks	launch tasks via cluster manager	execute tasks
	submit each stage as ready	retry failed or straggling tasks	store & serve blocks