Institute of Information Technology (IIT)

Jahangirnagar University

Course Code: MICT 5304

Course Title: Big Data Analytics

<u>Assignment - 01</u>

Submitted to:

Md. Fazlul Karim Patwary

Professor

IIT, JU

Submitted by:

Name: Md. Shakil Hossain

Roll No: 1061

MSc 2nd Semester

IIT, JU

Submission Date: 05/08/2025

What are the differences between Hadoop and Spark?

Feature	Hadoop	Spark
Processing Model	Disk-based, MapReduce batch jobs	In-memory, DAG-based, supports RDDs
Performance	Slower, high latency	Much faster, low latency
Real-time Support	No, batch processing only	Yes, real-time streaming and analytics
Ease of Use	Complex, low-level APIs	Easy, high-level APIs in multiple languages
Cost & Scalability	More affordable, easily scalable	Higher memory cost, harder to scale
Storage	Native HDFS	Uses HDFS, S3, other storage
Machine Learning	Requires external tools	Built-in MLlib library
Fault Tolerance	HDFS replication	RDD lineage and precomputation

How does PageRank determine the importance of a web page?

PageRank determines the importance of a web page by analyzing the quantity and quality of links pointing to it. Each link to a page is treated as a "vote" for that page's relevance or authority. However, not all votes are equal links from highly authoritative pages have much greater impact than those from less significant or low-quality pages.

Here's how PageRank establishes importance:

- **Link Analysis:** Every web page starts with an initial value. The algorithm examines all inbound links to a page, considering both the number and the authority of the linking sites.
- **Weight Distribution:** When a page links to several other pages, it distributes its own PageRank value among those pages. Thus, receiving a link from a highly ranked page passes more importance than from a low-ranked one.
- **Iterative Calculation:** The process is repeated over many iterations. At each step, a page’s new ranking is recalculated based on the rankings of the pages linking to it, until the system stabilizes.
- **Damping Factor:** To avoid infinite cycles and to make the system more realistic, a damping factor (often set to 0.85) is used. This reflects the probability that a user randomly jumping through links might at some point land on a completely different page not linked from the previous one.

Why is high-speed search important in systems like Google?

High-speed search is important in systems like Google because it significantly enhances user experience by delivering fast, relevant search results without delays. Users expect quick responses when they enter queries, and faster search results increase satisfaction and engagement. Google's emphasis on speed reflects its goal to serve content efficiently, as slow-loading pages lead to user frustration and increased bounce rates, reducing overall traffic and loyalty.

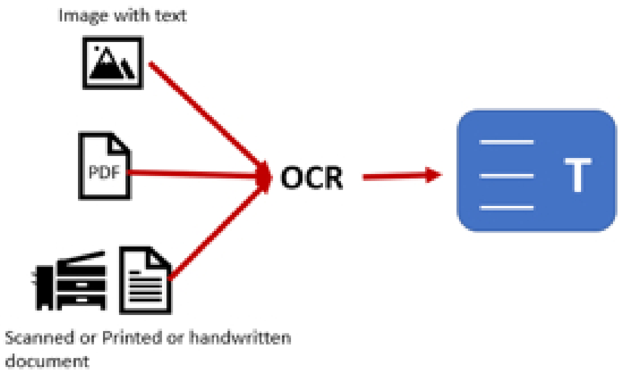
Moreover, speed affects Google's search ranking algorithms. Google considers page speed a ranking factor, rewarding faster-loading websites with better visibility in search results. This is because fast websites improve crawl efficiency for Google's bots and encourage users to stay longer, increasing conversions and site authority.

In summary, high-speed search improves:

- User satisfaction and engagement by reducing wait times.
- Search rankings since Google favors faster sites.
- Conversion rates and potential revenue.
- Crawl efficiency and the ability to index more content effectively.

What is OCR and how is it used in image analysis?

Optical Character Recognition (OCR) is a technology that converts images containing text such as scanned documents, photos of printed pages, or handwritten notes into machine-readable and editable text. Widely used in both business and research, OCR streamlines the process of digitizing printed materials, making information accessible for editing, searching, and automated data processing.



How OCR Works

OCR uses a combination of image analysis, pattern recognition, and advanced machine learning techniques to recognize text:

- **Image Preprocessing:** The input image is cleaned to remove noise, correct skew, enhance contrast, and often binarized so text stands out from backgrounds. This stage is crucial for improving recognition accuracy, especially with poor-quality images.
- **Text Detection & Recognition:** OCR systems segment the image into regions, lines, words, and characters. They then analyze the shapes of these regions using pattern matching or feature extraction to identify individual characters. Modern OCR engines use deep learning to improve recognition, even for varied fonts and complex backgrounds.
- **Postprocessing:** Recognized text is compiled, checked for errors using dictionaries or language models, and output as structured, machine-encoded text (such as in TXT, DOCX, or searchable PDF formats).

Applications in Image Analysis

OCR is a core component of image analysis wherever textual information needs to be extracted from images, including:

- **Digitizing Printed Documents:** Automates the conversion of books, invoices, historical documents, and forms into searchable and editable digital formats.
- **Scene Text Recognition:** Extracts text from photographs taken in everyday settings, such as street signs, license plates, storefronts, or product labels, enabling tasks like automated translation, navigation, and augmented reality overlays.
- **Handwriting Recognition:** Advanced OCR can transcribe handwritten notes or cursive writing, making it easier to archive and process analog records.
- **Business Process Automation:** OCR is used in sectors like banking, healthcare, postal services, and many more for efficient data entry and workflow automation.

How can big data technologies support image and OCR analysis?

Big data technologies play a crucial role in supporting image and Optical Character Recognition (OCR) analysis by enabling the processing, storage, and analysis of massive and diverse datasets, often in real time. Here’s how these technologies empower OCR and image analysis:

1. Scalable Data Storage and Processing

- **Handling Volume and Variety:** Big data frameworks such as Hadoop and Apache Spark store and process enormous quantities of images and documents, which is essential for modern OCR systems that process millions of scanned files, photos, and other visual data sources.

- **Efficient Batch and Stream Processing:** With Spark, for example, organizations can process images and extracted text quickly either in batches or streams.

2. Automated Extraction and Organization

- **OCR as a Front-End Pipeline:** OCR engines digitize text from images, transforming unstructured visual data into structured, machine-readable content. Big data platforms then ingest, index, and organize this output for further analysis and immediate access.
- **Enabling Search and Retrieval:** Once digitized and indexed in a big data environment, image-derived text becomes searchable, supporting intelligent document management, compliance checks, and audit trails across industries.

3. Advanced Analytics and Insights

- **Pattern, Trend, and Sentiment Analysis:** After OCR transforms images into data, big data analytics tools (using AI, machine learning, and deep learning) mine this information for patterns—such as fraud detection, customer trends, anomaly identification, or operational inefficiencies.
- **Predictive & Visual Analytics:** Big data analytics can forecast trends based on image or text data (like recognizing prescription patterns in healthcare, or sentiment in social media images), and generate visual dashboards for business intelligence.

4. Automation and Workflow Optimization

- **Workflow Automation:** Integrated OCR and big data solutions automate manual tasks for faster throughput and reduced errors, freeing human resources for higher-level work.
- **Real-Time Processing:** In sectors like logistics or finance, data extracted from images or scanned documents can be processed instantly to update records, alert decision-makers, or trigger downstream workflows and notifications.

5. AI-Enhanced Recognition and Continuous Learning

- **Leveraging Machine Learning at Scale:** Modern OCR tools use big data to train and refine AI models, improving recognition accuracy for diverse scripts, languages, or low-quality images. Large-scale data processing is crucial for keeping AI-OCR effective across ever-growing data sets and use cases.
- **Continuous Model Improvement:** Big data platforms enable continuous feedback, retraining, and improvement of OCR engines based on new image data and corrections from users.

6. Practical Applications Across Industries

- **Healthcare:** Analyzing medical images, records, and doctor’s handwritten notes for better diagnosis and compliance.
- **Finance:** Automating extraction from checks, invoices, and receipts; supporting fraud detection and auditing.
- **Legal, Retail, Logistics:** Instant processing of contracts, archiving documents, or tracking packages streamlining operations at scale.

Give one example of a real-world OCR application in big data analytics.

A prominent real-world example of OCR in big data analytics is automated data extraction from receipts and invoices in the finance industry. Companies use OCR to scan and digitize large volumes of receipts and invoices, converting unstructured paper or image data into structured formats. This automation enables seamless expense reporting, financial analysis, and compliance checks by integrating extracted data into big data systems, minimizing manual entry errors, and accelerating business workflows. For instance, financial assistant apps often rely on OCR for real-time receipt recognition and categorization, allowing for advanced analytics and insights on company-wide spending patterns.