1. Introduction (LATER)

2. Big Data Overview

 2.1 Definition and Characteristics

Big data is a widely used term today, yet its origins remain somewhat unclear. It likely emerged from discussions at Silicon Graphics Inc. in the mid-1990s. However, the term gained momentum much later, largely due to the explosive growth of digital data generated by mobile devices, social media platforms, sensors, and online transactions. As organizations began to recognize the potential value hidden in massive and complex datasets, the need for specialized tools and techniques to store, manage, and analyze this information became evident. This growing demand helped propel "big data" into mainstream discourse, making it a central concept in both industry and academia.

Understanding big data involves acknowledging its fluid and evolving nature across multiple dimensions. Initially, organizations like Gartner, Inc. and the TechAmerica Foundation defined big data as high-volume, complex, and fast-moving information requiring advanced processing techniques. Gartner introduced the foundational “Three Vs”: volume, velocity, and variety. This framework was later expanded, first to 5Vs with the addition of veracity and value, and then to 7Vs by including visualization and variability. Some sources have proposed even more elaborate frameworks, citing 10Vs or, in extreme cases, as many as 51. However, the 7Vs are widely accepted and provide a comprehensive enough foundation to understand the key characteristics and challenges associated with big data. These dimensions are interdependent and evolve with time, technology, and industry standards.

The 7Vs of Big Data:

1. Volume – Size

Refers to the massive amount of data generated from various sources (e.g., sensors, social media, transactions). Big data is typically measured in terabytes or petabytes. Its scale challenges storage, processing, and management systems.

2. Variety – Diversity

Describes the structural diversity of data, including structured (tables), semi-structured (XML), and unstructured (text, images, video). Greater variety increases complexity but also enhances insights. Advances in analytics have enabled firms to leverage previously untapped data types for strategic decision-making.

3. Velocity – Speed

This rapid generation requires fast processing to enable timely decisions—such as real-time fraud detection, personalized marketing, or dynamic pricing.

Denotes the rate of data generation and processing. Applications like fraud detection, personalized marketing, or real-time pricing depend on real-time or near-real-time data analysis.

4. Veracity – Trustworthiness

Focuses on the quality, accuracy, and reliability of data. Inconsistent, incomplete, or misleading data must be addressed to ensure valid analysis. Ensuring accuracy involves validating data sources and applying tools to manage uncertainty. High-veracity data leads to more reliable insights and decisions.

5. Value – Usefulness

Represents the potential of big data to deliver actionable insights and tangible business benefits. The goal is to extract valuable knowledge that supports better innovation, decisions and outcomes.

6. Variability – Inconsistency

Refers to the changing nature and flow of data over time. Fluctuations in structure, meaning, or speed require adaptive tools and flexible data handling methods.

7. Visualization – Clarity

Involves transforming data in intuitive, interactive formats like charts or dashboards to help users quickly understand trends, patterns, and relationships. Effective visualization helps stakeholders grasp trends, patterns, and anomalies that might otherwise remain hidden in raw data.

(…from the 51Vs paper, and )

 2.2 Technologies and Tools

Should I compare spark with Hadoop mapreduce here? Or just stick with Spark-next section and how long should this section be?

3. Analysis Techniques

 3.1 Apache Spark and Spark SQL

 3.2 Best Practices

4. OpenAlex Dataset

 4.1 Dataset Overview

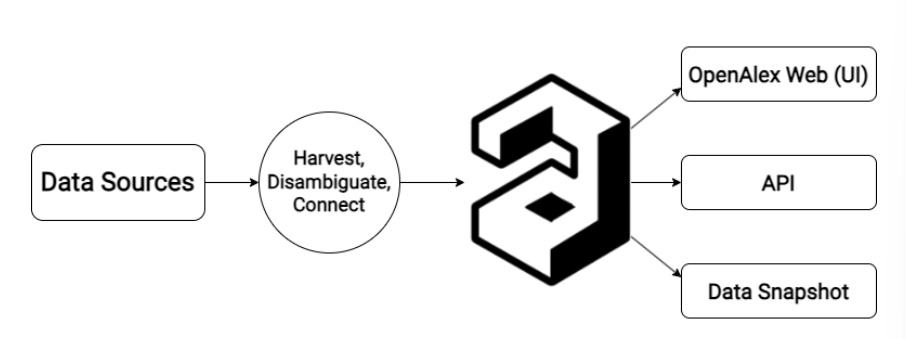
OpenAlex is a comprehensive catalog of scholarly entities and their relationships. These entities include works (e.g., articles, datasets, books), authors, sources, institutions, topics, publishers, and funders……. Together, they form a massive, interconnected web, technically a heterogeneous directed graph, containing hundreds of millions of entities and billions of connections.

It not only tracks metadata like titles and creation dates, but also disambiguates and links entities through citations, authorship, institutional affiliations, topics, and more. This structure enables powerful large-scale analysis and exploration of the research landscape.

It gathers and standardizes data from many sources, primarily **MAG** and **Crossref**, along with others like **ORCID**, **ROR**, **DOAJ**, **Unpaywall**, **PubMed**, **PubMed Central**, **ISSN Centre**, **Internet Archive**, web crawls, and repositories such as **arXiv** and **Zenodo**.

Access Methods:  
OpenAlex offers its data and analytics through three main channels:

* **OpenAlex Web** — A user-friendly web interface for exploring the data
* **OpenAlex API** — A fast, modern REST API for programmatic access
* **Data Snapshot** — A downloadable, free, periodic snapshot of the entire dataset



As of today, OpenAlex contains over 250 million scholarly works from 250,000 sources, linked to 90 million disambiguated authors from 100,000 institutions. The dataset includes broad coverage of the humanities, non-English languages, and research from the Global South. It is enriched with topic classifications, SDG tags, citation metrics, and more. For our analysis, we will use a complete snapshot of the OpenAlex dataset, totaling approximately 3.5 terabytes of structured data.

 4.2 JSON Data Handling

All snapshot data is stored in **Amazon S3** and is publicly accessible here: <https://openalex.s3.amazonaws.com/browse.html>

The dataset is organized into **seven JSON Lines files**, one for each entity type: works, authors, sources, institutions, topics, publishers, funders. Each line is a single JSON object, making the format easy to parse at scale. The entire snapshot is updated monthly.

The OpenAlex dataset exemplifies the 7Vs of big data, positioning it as a robust resource for large-scale computational analysis. It offers substantial **volume**, encompassing approximately 3.5 terabytes of structured scholarly metadata. Its **velocity** is reflected in frequent incremental updates and continuous ingestion pipelines aggregating data from over 250,000 diverse sources. The dataset exhibits significant **variety**, covering multiple academic disciplines, languages, and data types, including comprehensive metadata on topical classifications, citation networks, and Sustainable Development Goals (SDGs). Through rigorous author and institution disambiguation algorithms and transparent data provenance mechanisms, OpenAlex ensures high **veracity** and data reliability. Addressing **variability**, the dataset’s rich, multi-dimensional schema supports complex querying and multi-faceted analyses across temporal, geographic, and thematic dimensions. The **value** of OpenAlex derives from its open-access policy, extensive global scope, and enriched annotations, enabling advanced bibliometric, scientometric, and research evaluation applications. Finally, the dataset’s standardized structure and compatibility with contemporary data science ecosystems facilitate effective **visualization** and insight generation. Collectively, OpenAlex provides a scalable, high-quality foundation for data-driven scholarship in the science of science and research analytics domains.

5. System and Methods

 5.1 System Setup

What should I mention here?

 5.2 Data Processing Workflow

6. Data Analysis

 6.1 Collaboration Networks

I guess how are the entities connected, how will that be of value for our analysis

 6.2 Topic and Citation Trends – maybe we can rename this?

7. Discussion (LATER)

8. Conclusion (LATER)

9. References

For now just here (Armbrust, Xin et al. 2015) (Priem, Piwowar et al. 2022) (Khan, Naim et al. 2019) (Team 2025)

Armbrust, M., et al. (2015). Spark SQL: Relational Data Processing in Spark. Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data. Melbourne, Victoria, Australia, Association for Computing Machinery**:** 1383–1394.

Khan, N., et al. (2019). The 51 V's Of Big Data: Survey, Technologies, Characteristics, Opportunities, Issues and Challenges. Proceedings of the International Conference on Omni-Layer Intelligent Systems. Crete, Greece, Association for Computing Machinery**:** 19–24.

Priem, J., et al. (2022). "OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts." arXiv preprint arXiv:2205.01833.

Team, O. (2025). "OpenAlex Technical Documentation." from <https://docs.openalex.org/>.