Rajat--

Slide 1: Introduction to Pushshift Reddit Dataset

Presenter Notes:

- “Let’s start with a simple question: What is Pushshift? At its core, Pushshift is a large-scale social media dataset that includes all Reddit submissions and comments from 2005 to the present, updated in real-time. This dataset is a game-changer for social media research.”

--Change---

- “Why Reddit, you might ask? Reddit is unique because it has millions of subreddits, billions of comments, and relatively open data access, especially compared to platforms like Facebook and Twitter. This openness makes it a rich source for studying social behavior, online trends, and even political discourse. Researchers can dive deep into diverse communities—from niche hobbies to large-scale social movements.”

--Change---

- “The goal of Pushshift is simple yet powerful: to make Reddit data easily accessible and ready to analyze. Instead of spending hours—or even days—collecting, cleaning, and storing data, researchers can focus on the actual analysis. Pushshift offers APIs and monthly data dumps to streamline this entire process.”
- **Engage the audience:** "How many of you have ever tried scraping data from social media for research? It's time-consuming, right? Pushshift changes that by giving you clean, ready-to-use data."

Slide 2: Dataset Composition

Presenter Notes:

- “Now, let's talk numbers—this dataset is massive. Between June 2005 and April 2019, Pushshift gathered **651 million submissions** and **5.6 billion comments** from over **2.8 million subreddits**. That’s a staggering amount of data, and it grows every day.”
- "But it's not just about the size. The dataset contains both metadata—things like post IDs, timestamps, and authors—and the actual content, such as comment text and submission titles. This allows researchers to not only analyze what’s being said but also when and where conversations are happening."

  --Change--

- “The data is stored in a newline-delimited JSON format, or **ndjson**, which makes it efficient and easy to process programmatically. Each entry in the dataset includes detailed fields like **score**, **author**, **subreddit**, and even whether a comment is considered controversial or has been edited. This granular level of detail allows for sophisticated analyses.”
- **Engage the audience:** “If you were given access to 5.6 billion comments, what would you research? What trends or behaviors would you want to study?”

---Change--

Here is the Submission Json Response

---Change--

Here is the Comment Json Response

Slide: Conclusion

Presenter Notes:

- “So, what did we achieve with the **Pushshift Reddit Dataset**? This dataset gives researchers access to Reddit’s vast data, including **651 million submissions** and **5.6 billion comments** collected since 2005.”
- “Why is this so valuable? Pushshift removes the need to handle data collection and cleaning, giving you access to **APIs and monthly data dumps** that allow researchers to start their analysis faster.”
- “It also provides additional tools like the **Slackbot** for real-time **data interaction and visualization**, making collaboration much easier—something very few datasets offer.”
- *Ask the audience:* “Who wouldn’t want to skip all the tedious work of collecting and cleaning data and jump straight to analysis?”
- “With over **100 peer-reviewed publications** already using this dataset, it has shown its value across diverse fields—from **social science** to **disinformation research**.”
- “Looking ahead, as the dataset grows, Pushshift will continue to be an essential tool for exploring online behaviors and **community governance**.”

Sanjna---

Slide 3: Let's navigate the Post-API Landscape:

1) First we have Challenges in the Post-API age:

   There is restriction of data access by major platforms like Facebook and Twitter that is know as Post-API age.

   Because of the privacy concerns, researchers face difficulty in gathering social media data specifically for studying issues like online misinformation.

   With the restrictions of APIs, collecting data takes time and has become very difficult.

2) To overcome these challenges the solution is pushshift so lets discuss the role of pushshift.

   Unlike platforms where there is limited data access, pushshift provides large dataset of reddit posts and comments for both real time and historical data.

   This API allows researchers to access large volume of information without downloading.

   It reduces the storage, and makes data more available to wide range of users.

3) Why Pushshifts's API over reddit API?
   Because of its enhanced API functionality.

   It allows full text search against comments and submissions, also has larger single query limit.

   It handles five times more data per query than reddit's 100 object limit. Also it offers aggregation endpoints to provide summary. In this reddit lacks entirely.

   It helps researchers to focus on actual analysis rather than spending time on data collection, cleaning and storage.

Slide 4: Pushshift uses multiple backend software components to collect, store, catalog, index, and disseminate data to endusers:

1) Ingest Engine:
   It is responsible for collecting and storing raw data. It can be thought of as a framework for large scale collection of social media data sources.
   It also provides and manages job scheduling queue. It uses Redis queue for collected data before processing by any custom scripts.

2) Postgres:
   It is being used during advanced query for collected data, in that way it helps in effecient and structured data retrieval .

3) Elastic search: .

Pushshift currently usesElasticsearch (ES) as a scalable document store for each data source that is part of the ingest pipeline.

To handle reddit evolving api, it supports full text search and dynamic mapping. It also indexes and aggregates data.

The important feature for storing and analyzing data is scaling by utilizing cluster approach. It also supports Unicode and complete emoji search.

Here is the diagram that explain the Pushshift's Reddit data collection platform:

1. **Ingest Engine:** The platform ingests Reddit data through an ingest engine that collects posts and comments from Reddit's ecosystem, feeding the data into the processing pipeline.
2. **Storage and Queryability:** The data is stored in queryable databases like PostgreSQL and Elasticsearch, where Elasticsearch provides indexing and search functionalities to optimize data retrieval.
3. **Access via API and Archives:** The collected and stored data is made accessible through an API for users to query, and archived data is stored for long-term access and historical analysis.

Shivani Script---

Slide 5:

[Opening for Slide] Let's look at the tangible impact of the Pushshift Reddit Dataset in the academic community over the last few years.

[Details on Publication Trends] This bar graph represents the number of research papers published using Pushshift data from 2016 to 2019. Observing the trend, we can see a clear and significant increase in its utilization over these years:

- **2016**: The usage starts modestly, reflecting the initial stages of researchers exploring the potential of the Pushshift dataset.
- **2017**: There is a noticeable uptick in publications. This increase suggests that as researchers began to realize the value of this comprehensive dataset, its adoption in various studies grew.
- **2018**: The upward trend continues, indicating a solidifying trust in the dataset's reliability and relevance for social media research.
- **2019**: We see a substantial surge in the number of publications. This spike underscores a widespread recognition and integration of Pushshift data into mainstream social media research, marking it as an essential resource for academics.

[Closing for Slide] The escalating number of publications not only validates the dataset's utility but also highlights its growing importance in the field of web and social media analysis. Pushshift has evidently become a cornerstone resource for researchers looking to understand the complexities of online social interactions.

[Transition to Next Topic] With this understanding of Pushshift's academic influence, let's now delve into specific case studies where this dataset has enabled breakthrough insights in areas like community governance and misinformation.

Online Community Governance

"In the realm of Online Community Governance, the Pushshift dataset provides invaluable insights into the decentralized moderation systems of platforms like Reddit. This data allows researchers to analyze how volunteer moderators influence community norms and manage content, offering a unique contrast to the centralized models of other social media giants. Studies using Pushshift have helped us understand which moderation practices foster positive user engagement and which may lead to controversy or dissent."

Understanding Online Extremism

"Moving to Understanding Online Extremism, the dataset serves as a critical tool for researchers studying the dynamics of radicalization in online spaces. By providing access to discussions from various subreddits, including those frequented by extremist groups, Pushshift has enabled studies that track the migration and evolution of extremist ideologies. This is crucial for developing strategies to combat radicalization and understanding the effectiveness of platform policies against such activities."

Tackling Online Disinformation

"In the battle against Online Disinformation, Pushshift has been instrumental. Researchers utilize this dataset to trace the origins and spread of misinformation campaigns, especially during pivotal events like elections. The longitudinal data allows for an analysis of how narratives evolve over time, offering insights into the mechanisms of misinformation spread and the role of social media in amplifying these narratives."

Advancing Web Science

"In Advancing Web Science, Pushshift supports studies at the intersection of technology and society. Researchers explore how design choices on social media platforms impact user behavior and community success. The data aids in understanding the spread of technological innovations and the dynamics of online communities, providing empirical evidence to guide the development of more effective online environments."

Enabling Big Data Science

"For Enabling Big Data Science, the dataset is a cornerstone for researchers developing new algorithms and models capable of handling large-scale social media data. Pushshift's extensive repository of text data facilitates foundational research in network science, helping to uncover patterns and relationships within social media interactions."

Innovations in Health Informatics

"In the field of Health Informatics, the relative anonymity of Reddit captured in Pushshift's data allows for candid discussions on sensitive health topics. This is invaluable for researchers studying phenomena like mental health issues, addiction, and disease communities, offering insights into patient behaviors and treatment discussions that are often challenging to access through traditional health data sources."

Empowering Intelligent Systems

"Lastly, in Empowering Intelligent Systems, Pushshift's dataset accelerates the development of sophisticated machine learning models, particularly in natural language processing. By analyzing vast amounts of human-generated text, researchers can improve conversational agents, enhance sentiment analysis tools, and contribute to the growing field of AI, making systems more responsive and understanding of human language."

Slide 6:

[Introduction to Data Collection Services] Traditional data storage models, often limited to static "buckets" of data hosted on cloud servers, are being transformed. Modern services are not just repositories; they are dynamic platforms providing real-time, large-scale data alongside powerful tools for analysis and interaction.

[Pushshift's Unique Position] Enter Pushshift—a service that exemplifies this new paradigm. While there are many players in the field of social media data collection, Pushshift uniquely focuses on Reddit, offering an extensive dataset that includes historical and real-time data. What sets Pushshift apart is its dedication to making Reddit data easily accessible and highly usable for researchers, thereby reducing the time from data acquisition to insight generation.

[Comparison with Key Alternates] To contextualize Pushshift's contributions, let's consider some key alternate services:

4. **Media Cloud:** This platform tracks millions of news stories, providing aggregated data through a semi-public API, aiding researchers in media studies.
5. **GDELT:** Monitoring global news media, GDELT offers insights into worldwide events and trends, serving as a comprehensive resource for geopolitical research.
6. **Stats Exchange:** Known for its roots in Stack Overflow, this service allows complex SQL queries across its social Q&A data, enriching research in technical and community-driven discussions.
7. **Wikimedia:** As the backbone of Wikipedia, it provides vast data dumps and APIs, supplemented by interactive tools like Jupyter Notebooks, facilitating diverse research from historical revisions to traffic analysis.

[Feature Comparison] Now, let's dive deeper with a feature comparison:

- All services, including Pushshift, offer public APIs and regular data updates.
- Where Pushshift shines is its integration of interactive computing and community engagement, although these are areas still under active development compared to some more established platforms.
- Pushshift’s commitment to outreach and community-building is growing, aiming to enhance user support and engagement over time.

[Closing] In conclusion, while many services offer valuable data, Pushshift stands out for its focus on Reddit and its effort to lower barriers for researchers studying social media dynamics.