BIG DATA PROCESS MAPPING

Prepared by :PRAGATHI S

USN:01SU24CS104

Date:23-02-2026

1.Introduction

Big data systems have become integral to modern businesses, enabling organizations to collect, store, and analyze massive volumes of structured and unstructured data in real time. These systems provide insights that drive personalized services, improve operational efficiency, and support strategic decision-making.

The purpose of this report is to map the data flow of a real-world big data system—Amazon Recommendations. This includes identifying the data sources, storage mechanisms, processing techniques, and outputs. By examining how Amazon processes vast amounts of user and product data to generate personalized recommendations, the report highlights the architecture, technologies, and workflows involved in a high-scale big data application.

This study provides an understanding of how big data systems operate in practice and demonstrates the importance of data integration, analytics, and machine learning in delivering real-time, user-centric services.

2. Objectives

The main objectives of this report are:

1. **To analyze a real-world big data system** – Understanding how Amazon Recommendations collects, stores, and processes data.
2. **To map the data flow** – From data sources to storage, processing, and final output.
3. **To identify key technologies and methods** – Highlighting tools, databases, and machine learning techniques used in big data systems.
4. **To evaluate system efficiency and scalability** – Understanding how the system handles large volumes of data in real time.
5. **To provide insights into practical applications of big data** – Demonstrating how data-driven decisions improve user experience and business outcomes.

3.Data Sources

The Amazon Recommendation System relies on **large volumes of diverse data** collected from multiple sources. These data sources are crucial because the quality and variety of input data directly affect the accuracy of recommendations. They can be grouped as follows:

3.1 User Behavior Data

This includes all interactions users have with the Amazon platform:

- **Clicks and searches:** Every time a user clicks a product or performs a search, the system records this activity.
- **Browsing patterns:** The system tracks which categories, products, or brands a user frequently visits.
- **Dwell time:** Time spent on each product page indicates interest or purchase intent.
- **Navigation flow:** How users move from one page to another helps the system predict preferences.

Example: If a user frequently searches for "wireless headphones" and spends time reading reviews, the system notes this interest to recommend similar or higher-rated headphones.

3.2 Purchase History

- **Transaction records:** Products bought, quantity, and purchase frequency.
- **Shopping cart and wishlist data:** Items added to carts or wishlists indicate potential future purchases.
- **Returns and cancellations:** Helps the system learn about products users dislike or avoid.

Example: If a user buys cooking books frequently, Amazon may recommend kitchen gadgets or recipe tools based on this behavior.

3.3 Ratings and Reviews

- **Star ratings:** Quantitative feedback on user satisfaction.
- **Text reviews:** Qualitative data providing insights into user preferences.
- **Engagement with reviews:** Votes on helpful reviews or comments can influence recommendation weight.

Example: A product with many 5-star reviews similar to those a user liked may appear higher in recommendations.

3.4 Product Metadata

- **Attributes:** Product category, brand, price, color, size, and specifications.
- **Availability and pricing updates:** Information on discounts, stock, and promotions.

- **Related products:** Bundled items or complementary products for cross-selling opportunities.

Example: If a user buys a camera, the system may suggest compatible lenses or tripods using product metadata.

3.5 Third-Party and External Data

- **Social media trends:** Mentions, likes, and shares of products can inform trending items.
- **Market data:** Competitor pricing, seasonal trends, or product popularity across platforms.
- **External reviews:** Ratings from partner websites may enhance recommendation accuracy.

3.6 Real-Time Event Data

- **Live actions:** Real-time clicks, searches, or purchases immediately influence the recommendation engine.
- **Dynamic personalization:** The system updates suggestions as the user interacts with the platform, creating a responsive experience.

Example: If a user searches for "laptop bags" today, the homepage may immediately show bags even if the user has not purchased them before.

4.Data Storage

After data is collected from multiple sources, it needs to be **stored efficiently** to support processing, analytics, and real-time recommendations. Amazon uses a combination of storage systems to manage the massive volume, variety, and velocity of its data.

4.1 Data Lakes

- **Purpose:** Store raw, unstructured, semi-structured, and structured data at scale.
- **Technology Example:** Amazon S3 (Simple Storage Service)
- **Function:** All incoming data—clickstreams, reviews, product metadata—is stored in its raw form for future processing and analysis.
- **Benefit:** Flexible storage allows for different data formats without strict schema requirements.

4.2 Data Warehouses

- **Purpose:** Store structured data optimized for querying and reporting.
- **Technology Example:** Amazon Redshift

- **Function:** Aggregates structured data from data lakes for analytics, business intelligence, and reporting.
- **Benefit:** Supports complex queries efficiently, enabling analytics teams to track trends and patterns.

4.3 NoSQL Databases

- **Purpose:** Provide fast access to semi-structured or unstructured data.
- **Technology Example:** Amazon DynamoDB
- **Function:** Stores user profiles, session data, and product catalogs for low-latency access by the recommendation engine.
- **Benefit:** Ensures real-time responses for user interactions, supporting personalized recommendations instantly.

4.4 Caching Systems

- **Purpose:** Speed up access to frequently requested data.
- **Technology Example:** Redis or Memcached
- **Function:** Temporary storage of popular recommendations, user sessions, or trending products to reduce database load.
- **Benefit:** Reduces latency, ensuring users receive instant recommendations without delay.

4.5 Data Backup and Replication

- **Purpose:** Ensure data reliability and disaster recovery.
- **Function:** Critical datasets are replicated across multiple regions and servers.
- **Benefit:** Protects against data loss and ensures high availability for global users.

5.Data Processing

Once the data is collected and stored, Amazon's recommendation system processes it to extract insights, generate patterns, and produce personalized recommendations. The processing stage involves **batch processing, real-time stream processing, and machine learning workflows**.

5.1 Batch Processing

- **Purpose:** Analyze large volumes of historical data to identify patterns and trends.
- **Technology Examples:** Apache Hadoop, Apache Spark

- **Function:** Processes accumulated data such as past purchases, reviews, and browsing history to train recommendation models.
- **Benefit:** Enables the system to understand long-term user preferences and product correlations.

Example: Collaborative filtering models use historical purchase data to find patterns like “users who bought X also bought Y.”

5.2 Stream Processing

- **Purpose:** Handle real-time events for instant recommendations.
- **Technology Examples:** Amazon Kinesis, Apache Kafka, AWS Lambda
- **Function:** Processes live user activity, including clicks, searches, and product views, to update recommendations on the fly.
- **Benefit:** Ensures personalization is immediate, adapting to the user’s current behavior.

Example: If a user is browsing headphones right now, the system dynamically updates the homepage with relevant products.

5.3 Machine Learning Models

- **Collaborative Filtering:** Recommends products based on user-user or item-item similarities.
- **Content-Based Filtering:** Suggests products similar to those a user has interacted with based on product features.
- **Hybrid Models:** Combines collaborative and content-based filtering for better accuracy.
- **Deep Learning Models:** Uses neural networks to capture complex patterns in user behavior and product relationships.

Example: A neural network can predict which combination of products a user is likely to purchase together, improving cross-selling.

5.4 Feature Engineering

- **Purpose:** Transform raw data into meaningful inputs for machine learning models.
- **Examples of Features:** Purchase frequency, browsing time, product ratings, seasonal trends, and product categories.
- **Benefit:** High-quality features improve model accuracy and recommendation relevance.

5.5 Model Training and Evaluation

- Models are trained on historical data using batch processing.
- Regular evaluation ensures models remain accurate as user preferences and product catalogs change.
- Retraining is performed periodically or triggered by significant changes in data patterns.

6.Data Output

After processing, the insights generated by Amazon’s recommendation system are delivered through various outputs, both **to the end users** and **internally** for business intelligence. These outputs ensure that the system adds value in real-time and supports strategic decisions.

6.1 Personalized Recommendations for Users

- **Homepage suggestions:** Recommended products based on browsing and purchase history appear prominently on the homepage.
- **Product pages:** “Customers who bought this also bought…” or “Similar items” sections show relevant suggestions.
- **Emails and push notifications:** Personalized offers, deals, or reminders sent to users to drive engagement and repeat purchases.
- **Search result ranking:** Search results are reordered to prioritize products the system predicts the user is most likely to buy.

Example: A user who frequently buys fitness equipment may see suggestions for protein supplements, workout gear, or related fitness gadgets.

6.2 Real-Time Feedback Loop

- **User interactions with recommendations** are captured and fed back into the system.
- **Dynamic updates:** Real-time data ensures recommendations are adjusted instantly based on the user’s latest actions.
- **Continuous learning:** The feedback loop helps improve model accuracy over time.

Example: If a user clicks on a recommended item but does not purchase it, the system may adjust the weight of similar items in future suggestions.

6.3 Internal Business Insights

- **Analytics dashboards:** Product managers and marketing teams use data insights to track trends, measure recommendation performance, and make inventory or marketing decisions.
- **Performance monitoring:** Evaluates which algorithms or features produce the most effective recommendations.

- **Trend detection:** Identifies emerging products or customer interests to guide promotions and product launches.

7.Conclusion

Amazon's recommendation system demonstrates how big data can be effectively harnessed to provide personalized, real-time services to millions of users worldwide. By collecting data from diverse sources—user behavior, purchase history, ratings, product metadata, and external trends—the system builds a rich dataset that forms the foundation for analytics and machine learning.

The multi-tiered storage architecture, including data lakes, warehouses, NoSQL databases, and caching systems, ensures that this data is stored efficiently, securely, and is readily accessible for both batch and real-time processing. Advanced processing techniques, including collaborative filtering, content-based filtering, hybrid models, and deep learning, enable the system to generate accurate and relevant recommendations.

Finally, the outputs of the system—personalized suggestions, search result ranking, notifications, and internal analytics—demonstrate how big data adds value both for users and forbusiness decision-making. The case of Amazon Recommendations highlights that **effective big data systems rely on the seamless integration of data collection, storage, processing, and output** to deliver insights and improve user experiences.

8.References

1. Dean, J., & Ghemawat, S. (2008). *MapReduce: Simplified Data Processing on Large Clusters*. Communications of the ACM, 51(1), 107–113.

2. Marz, N., & Warren, J. (2015). *Big Data: Principles and Best Practices of Scalable Real-Time Data Systems*. Manning Publications.

3. Amazon Web Services. (2023). *Building Recommendation Systems on AWS*. Retrieved from https://aws.amazon.com/machine-learning/recommendations/

4. Aggarwal, C. C. (2016). *Recommender Systems: The Textbook*. Springer.

5. Chen, M., Mao, S., & Liu, Y. (2014). *Big Data: A Survey*. Mobile Networks and Applications, 19(2), 171–209.

6. Grolinger, K., Hayes, M., Higashino, W. A., L'Heureux, A., Allison, D., & Capretz, M. A. M. (2013). *Challenges for MapReduce in Big Data Applications*. Journal of Cloud Computing, 2(1), 1–21.

7. Bihani, P., & Patel, S. (2020). *Data Flow Architecture in Big Data Systems*. International Journal of Computer Applications, 176(33), 1–7.

8. Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., & Fuso Nerini, F. (2020). *The Role of Artificial Intelligence in Achieving the Sustainable Development Goals*. Nature Communications, 11, 233.

9. Zhang, S., Yao, L., Sun, A., & Tay, Y. (2019). *Deep Learning Based Recommender System: A Survey and New Perspectives*. ACM Computing Surveys, 52(1), 1–38.

10. Kaplan, J., & Haenlein, M. (2019). *Siri, Siri, in my Hand: Who's the Fairest in the Land? On the Interpretations, Illustrations, and Implications of Artificial Intelligence*. Business Horizons, 62(1), 15–25.