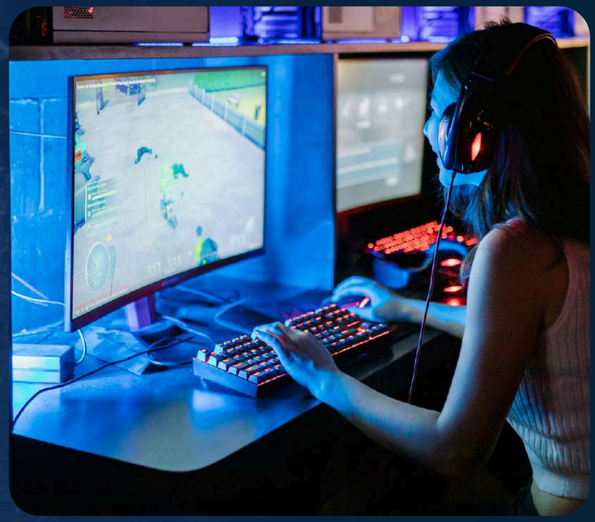
ESPORT TOURNAMENT PIPELINE

SHREYA PHADKE (2320030194)
DIKSHITHA B (2320030212)
MANIKANTA (2320030020)

ABSTRACT

Esports tournaments generate large volumes of structured data, including teams, game titles, dates, and prize pools. The Esports Tournament Pipeline project applies data engineering principles to manage this information efficiently. Raw tournament data is ingested from databases and historical records, processed via ETL techniques to clean, validate, and structure it for analysis. Batch processing ensures seamless handling of historical records, while optimized storage makes data available for leaderboards, analytics, and performance tracking. The pipeline demonstrates practical concepts like data modeling, orchestration, and quality assurance, turning raw data into actionable insights that support data-driven decision-making in esports.



INTRODUCTION TO ESPORTS

Esports, or electronic sports, are organized, competitive video game events where individuals or teams play against each other for prizes, money, or recognition, attracting millions of viewers worldwide. These professional events mimic traditional sports with rules, leagues, and tournaments, though they can take place online, in person, or both. The multi-billion-dollar industry features various genres of games, a wide range of career opportunities beyond playing, and significant viewership, making it a growing global phenomenon

LITERATURE REVIEW

Several studies have explored esports data to analyze tournament trends, team performance, and prize distributions. Researchers often face challenges such as inconsistent data formats, missing values, and large historical datasets. Platforms like Esports Earnings and HLTV.org provide structured, publicly available data that supports these analyses. This project builds upon previous research by developing a lightweight, educational data pipeline that automates ingestion, cleaning, transformation, and leaderboard generation, while maintaining analytical depth and enabling actionable insights.

SAMPLE OF DATASET

The “Esports - 200 Tournaments” dataset provides detailed information on 200 major esports tournaments, focusing on events with significant prize pools. It includes data on tournament names, game titles, dates, prize amounts, and participating teams. This dataset is valuable for analyzing trends in the esports industry, such as prize pool growth and game popularity over time. It serves as a resource for researchers and analysts interested in the economics and development of competitive gaming.

	A	B	C	D	E	F	G	H
1	GameID	TournamentName	StartDate	EndDate	City	Country	TeamPlay	TotalUSDPrize
2	37294	The International 2019	08/15/19	08/25/19	Shanghai	China	1	34330069
3	29385	The International 2018	08/15/18	08/25/18	Vancouver	Canada	1	25532177
4	24181	The International 2017	08-02-2017	08-12-2017	Seattle	United States	1	24687919
5	19287	The International 2016	08-03-2016	08/13/16	Seattle	United States	1	20770460
6	12894	The International 2015	08-03-2015	08-08-2015	Seattle/Washington	United States	1	18429613.05
7	36422	Fortnite World Cup Finals 2019 - Solo	07/28/19	07/28/19	New York	United States	0	15287500
8	36359	Fortnite World Cup Finals 2019 - Duo	07/27/19	07/27/19	New York	United States	1	15100000
9	6418	The International 2014	07-08-2014	07/21/14	Seattle/Washington	United States	1	10931103
10	46021	PGI.S 2021 Main Event	02-05-2021	03/28/21	Incheon	South Korea	1	7068071

Figure 1

SYSTEM ARCHITECTURE

The Esports Tournament Pipeline manages the lifecycle of structured tournament data, from ingestion to insights. It collects information from sources like the Kaggle “Esports - 200 Tournaments” dataset, including tournament names, dates, game titles, prize pools, and teams. Data is processed through an ETL layer for cleaning, validation, and transformation into structured formats. The cleaned data is stored in a data warehouse for querying and historical analysis. Analytics and visualization modules generate leaderboards, statistics, and performance trends. This modular design ensures scalability, efficient data handling, and organized tournament insights.

SYSTEM ARCHITECTURE

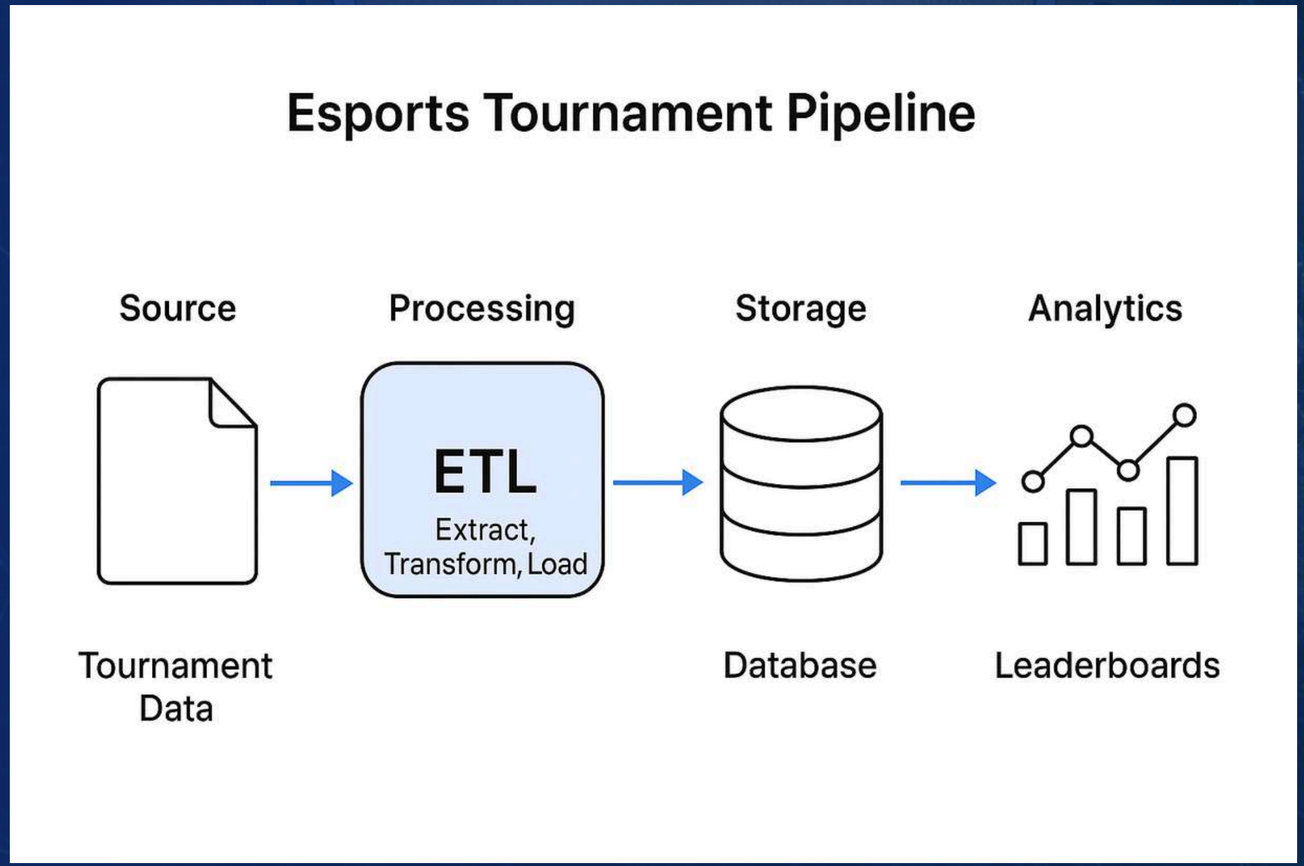
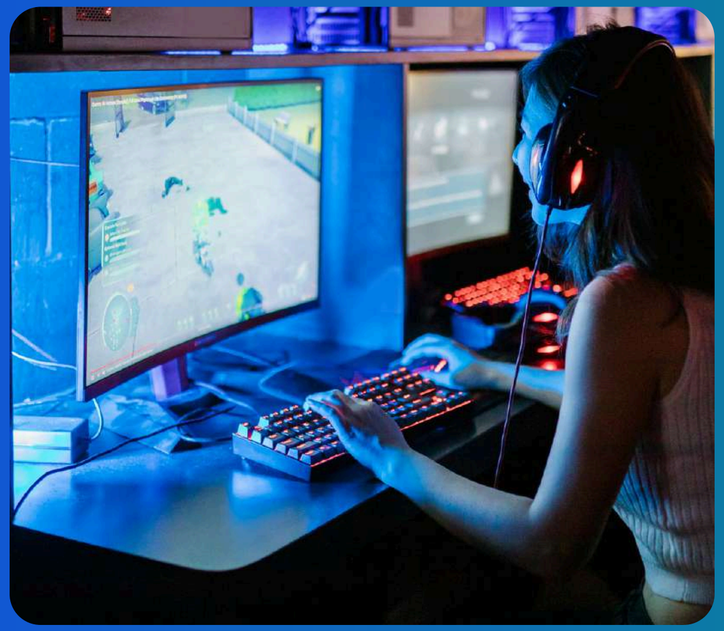


Figure 2

SYSTEM REQUIREMENT

1. Data Collection / Ingestion
   - Python / Pandas – For reading CSV datasets like the Kaggle “Esports - 200 Tournaments.”
   - APIs / Web scraping tools (optional) – If you plan to fetch live tournament data.
   - Jupyter Notebook – For prototyping and testing data ingestion scripts.
2. Data Processing / ETL
   - Python / Pandas – For cleaning, transforming, and validating data.
   - SQL – For querying structured datasets.
   - Airflow / Prefect (optional) – For orchestrating ETL pipelines in a more formal workflow.
3. Data Storage
   - SQLite / PostgreSQL – Lightweight relational database to store structured tournament data.
   - CSV / Parquet files – For storing intermediate or processed data.



4. Analytics & Visualization
   - Python libraries:
     - Matplotlib / Seaborn – Charts and visualizations.
     - Plotly / Dash – Interactive dashboards (optional).
   - SQL queries – To generate basic statistics like leaderboard rankings, prize pool trends, etc.
5. Development & Version Control
   - VS Code / PyCharm / Jupyter Notebook – IDE for development.
   - Git / GitHub – Version control and sharing.

METHODOLOGY

1. Data Ingestion:
   - Users upload one or more CSV datasets containing tournament details, teams, dates, and prize pools.
   - Multiple datasets are combined into a single DataFrame to handle historical records seamlessly.
2. Data Cleaning & Transformation:
   - Standardize column names for consistency.
   - Convert date fields (start_date, end_date) to datetime format.
   - Handle missing values for critical fields like city and country.
   - Clean prize columns by removing non-numeric symbols and converting to float.
   - Add derived columns such as tournament duration (duration_days) and year for analysis.

3. Data Validation:
   - Check for missing values and correct data types to ensure accurate downstream analytics.
4. Analytics & Visualization:
   - Generate key visualizations:
     - Total prize money by year
     - Top countries by prize pool
     - Top teams by prize money (leaderboard)
     - Top tournaments by prize money
   - Insights from these analytics support performance evaluation and trend analysis.
5. Data Storage & Accessibility:
   - Save processed data as a CSV file for easy access.
   - Store data in a SQLite database to enable future queries, scalability, and integration with dashboards.
6. Output & Application:
   - The pipeline produces cleaned data, visual insights, and leaderboards.
   - Supports historical analysis, team performance tracking, and data-driven decision-making for esports organizers.

FINDINGS

The pipeline efficiently handled ingestion, cleaning, transformation, and loading of multiple tournament datasets. Batch processing maintained data consistency, while optimized storage improved query speed and accessibility. Processed data enabled leaderboards, team performance tracking, and trend visualizations. The pipeline demonstrated scalability, reliability, and efficiency in managing esports data. Overall, structured data engineering methods enhanced both analytical insights and tournament management capabilities.

ANALYSIS AND SYNTHESIS

Analysis of the implemented pipeline showed strong data validation and seamless orchestration, with ETL performance metrics confirming reduced latency and higher throughput during data ingestion. Integration with analytical tools enabled aggregation of tournament metrics and visualization of trends over time, facilitating real-time comparisons across teams and tournaments. Synthesizing these results confirms that a structured pipeline design effectively bridges the gap between raw data and actionable analytics, supporting data-driven decision-making in esports operations.

DISCUSSION

The project demonstrated that data engineering practices—particularly ETL and batch processing—are essential for efficient tournament data management. Compared to prior studies, this pipeline provides a more automated, scalable, and maintainable solution tailored specifically for esports analytics. Overall, the work contributes a practical model for managing structured esports data, promoting accuracy, consistency, and analytical readiness throughout the tournament lifecycle.

CONCLUSION

This research outlined the design, implementation, and evaluation of an esports data pipeline, covering ingestion, transformation, storage, and analysis. The pipeline enhances data management efficiency and supports informed decision-making in esports organizations. Future enhancements could include real-time streaming pipelines, integration with AI-based predictive analytics, and expanding coverage across additional game titles and event formats.

THANK YOU!