**Business Intelligence Systems Assignment - Part 1: Sun Model and Star Schema Design**

**1. Introduction**

Web servers collect huge amounts of log data every day, but making sense of that data can be challenging without the right structure. This project tackles that problem by designing a **Business Intelligence (BI) system** for web log analysis. The system uses a **Sun Model** and **Star Schema** to transform raw log files into structured data that can be easily analysed and visualized.

The goal is to answer key business questions, like understanding user behaviour, tracking website performance, and identifying traffic sources.

**2. Sun Model Design**

At the heart of this system is the **WebRequests Fact Table**, which records each web request. Surrounding it are **five key dimensions** that provide context for deeper analysis. This structure, known as the **Sun Model**, makes it easy to answer important questions about user behaviour, traffic patterns, and website performance.

**Key Measures (Fact Table):**

* **BytesServed:** The size of each web response. Tracking this helps identify bandwidth-heavy pages.
* **ProcessingTime:** How long each request took to process, making it easier to spot slow-loading pages.

**Key Dimensions and Their Purpose:**

|  |  |  |
| --- | --- | --- |
| Dimension | Attributes | Why It Matters |
| Date | FullDate, Day, Month, Year | Helps analyse trends over time. |
| Client | IPAddress, City, Country | Shows where users are located and what devices they use. |
| Request | URL, FileType | Reveals which pages and file types are most popular. |
| Error | HTTPStatusCode, ErrorType | Identifies issues affecting user experience. |
| Referrer | ReferrerURL | Shows which websites are driving traffic. |

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

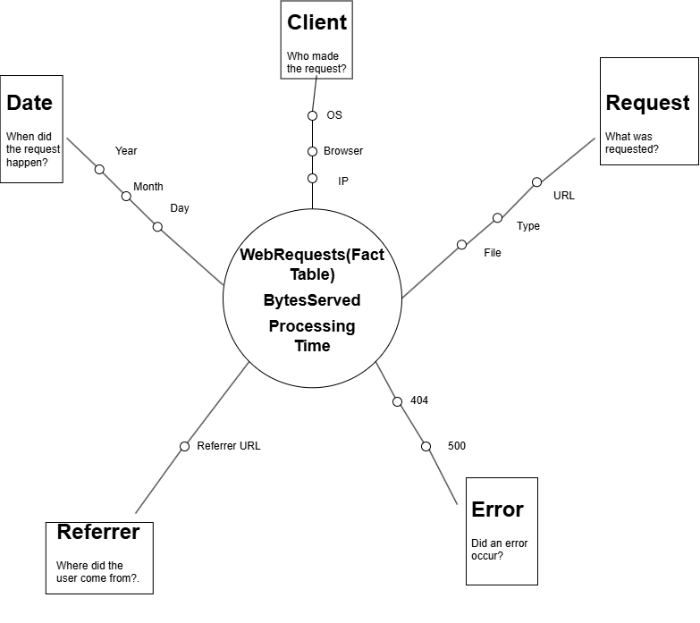


Figure 1: Sun Model

**3. Star Schema Implementation**

The **Star Schema** turns the Sun Model into a working database. The **WebRequests Fact Table** sits at the center, linked to each dimension table by foreign keys. This structure simplifies querying while keeping data organized and fast to retrieve.

**Fact Table:**

* **Primary Key:** (RequestID, RequestDate)
* **Foreign Keys:** DateID, ClientID, RequestDimID, ErrorID, ReferrerID
* **Measures:** BytesServed, ProcessingTime

**Dimension Tables:**

1. **DateDim:** Tracks request dates by day, month, quarter, and year.
2. **ClientDim:** Stores user IP addresses, cities, countries, browsers, and operating systems.
3. **RequestDim:** Records requested URLs and file types.
4. **ErrorDim:** Captures HTTP status codes and error descriptions.
5. **ReferrerDim:** Tracks traffic sources (e.g., search engines, social media).

This design ensures that queries run efficiently, even as the dataset grows.

**4. Expected Queries and Insights**

With the Star Schema in place, the following **queries** reveal valuable insights:

|  |  |  |
| --- | --- | --- |
| Business Question | Query Purpose | Business Insight |
| 1. Frequent Visitors | Who visits the site most often? | Identify loyal users and potential customers. |
| 2. Geo Trends | Where are users coming from? | Focus marketing efforts on high-traffic regions. |
| 3. Slow Pages | Which pages take the longest to load? | Prioritize performance improvements for better user experience. |
| 4. Error Monitoring | Are there spikes in errors like 404 or 500? | Quickly identify and fix issues before users are impacted. |
| 5. Bot Traffic | Which IPs repeatedly access robots.txt? | Filter bot traffic to ensure accurate reporting. |
| 6. Top Referrers | Which websites drive the most traffic? | Optimize marketing by focusing on high-performing channels. |
| 7. Traffic Patterns | Do users visit more on weekdays or weekends? | Schedule content releases based on peak traffic times. |
| 8. Errors by Device | Are errors more common on specific browsers? | Ensure cross-browser compatibility for a smoother experience. |

**5. Scalability and Performance**

To ensure the system can handle future growth, several **performance enhancements** were implemented:

1. **Partitioning:** The WebRequests table is partitioned by RequestDate, speeding up time-based queries.
2. **Indexing:** Non-clustered indexes were added to ClientID, ProcessingTime, and ReferrerID for faster lookups.

**6. Conclusion**

This project successfully transformed raw web logs into a structured, queryable system. Using the **Sun Model** and **Star Schema**, it became possible to uncover valuable insights into user behaviour, website performance, and traffic sources.

Thanks to **partitioning, indexing** the system is scalable and ready to handle larger datasets. With the **Power BI dashboard**, users will be able to explore the data visually and make informed decisions.