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

**1. Introduction**

This project develops a BI system for web log analysis using a **Sun Model** and **Star Schema** to support querying, visualization, and insights into user behavior, performance, and errors.

**Key Objectives:**

1. Create a **Sun Model** for facts and dimensions.
2. Build a **Star Schema** in SQL Server.
3. Define **expected queries** for insights.
4. Ensure **scalability** for big data.

**2. Sun Model Design**

The **Sun Model** represents web requests as **facts**, with **dimensions** providing context.

**Core Fact:** WebRequests – Unique web requests.

**Dimensions:**

1. **Date:** When the request occurred *(Day, Month, Year)*
2. **Client:** Request source *(IP, City, Browser, OS)*
3. **Request:** What was requested *(File type, URL)*
4. **Error:** Error occurrence *(404, 500, etc.)*
5. **Referrer:** Traffic source *(Referrer URL)*

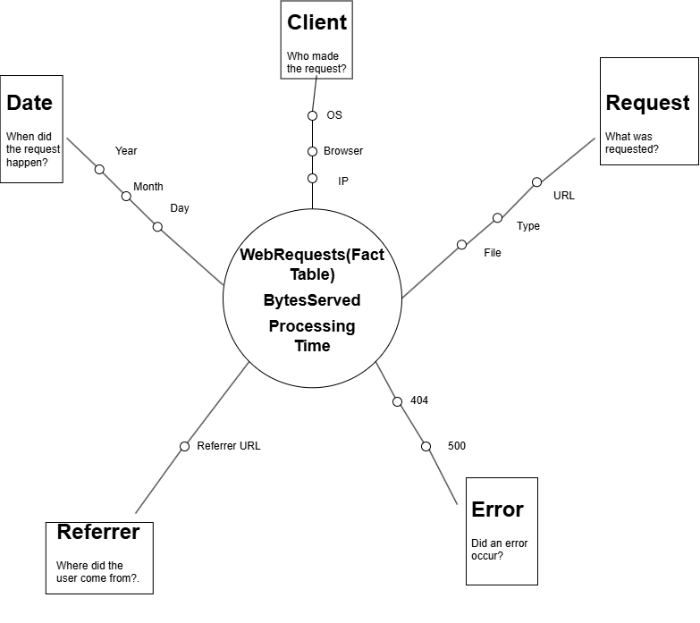


Figure 1: Sun Model

**3. Star Schema Design**

The **Star Schema** translates the Sun Model into one **Fact Table** and multiple **Dimension Tables**.

**Fact Table:** WebRequests

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

**Dimension Tables:**

1. **DateDim:** DateID (PK), FullDate, Day, Month, Year
2. **ClientDim:** ClientID (PK), IPAddress, City, Country, Browser, OS
3. **RequestDim:** RequestID (PK), FileType, URL
4. **ErrorDim:** ErrorID (PK), HTTPStatusCode, ErrorType
5. **ReferrerDim:** ReferrerID (PK), ReferrerURL

**4. Expected Queries and Insights**

| **Business Question** | **Query Description** | **Power BI Visualization** |
| --- | --- | --- |
| **Frequent Visitors** | Identify users with >5 visits. | Bar Chart (IP vs. Visits) |
| **GeoLocation of Users** | Identify user locations by IP. | Map Chart (City/Country) |
| **Slow-Loading Pages** | Top 5 pages by processing time. | Line Chart (URL vs. Time) |
| **Error Spike Detection** | Identify days with high 404/500 errors. | Pie Chart (Error Type) |
| **Bot Traffic Detection** | Detect IPs accessing robots.txt. | Column Chart (Bot Hits) |
| **Top Referrers** | Identify top 5 traffic sources. | Bar Chart (Referrer Count) |
| **Traffic Patterns** | Compare weekday vs. weekend traffic. | Column Chart (Day Type) |

**Example Insight:** *Frequent visitors can be identified by querying WebRequests and ClientDim for IPs with more than five visits.*

**5. Future Scalability for Big Data**

To handle larger datasets, the following enhancements are recommended:

1. **Partitioning & Indexing:**
   * Partition WebRequests by DateID for faster queries.
2. **Incremental Data Loading:**
   * Use **Change Data Capture (CDC)** to load only new logs.
3. **Power BI Optimization:**
   * Switch from **Direct Query** to **Import Mode** for faster reporting.

**6. Insights and Business Value**

This Star Schema delivers key insights:

1. **User Behavior:** Identify frequent visitors.
2. **Regional Trends:** Analyze user locations.
3. **Performance:** Spot slow-loading pages.
4. **Error Monitoring:** Track common errors (404, 500).
5. **Traffic Sources:** Identify top referrers.

**Example Insight:** *GeoLocation data highlights high-traffic regions for targeted marketing.*

**7. Conclusion**

The **Sun Model** and **Star Schema** efficiently organize web log data, enabling insights into user behavior, performance, and errors.

Future scalability is ensured through **partitioning**, **indexing**, and **incremental ETL**, maintaining system efficiency as data grows.

This design supports **Part 2**, focusing on ETL, Power BI visualizations, and advanced queries.