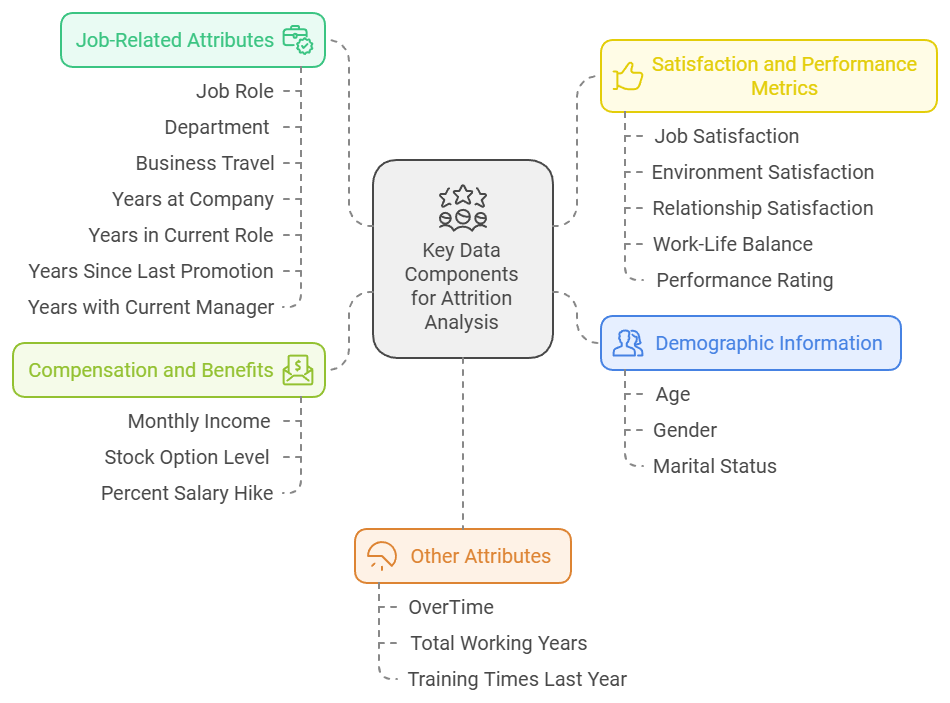
**Problem Statement:**

Identify Key Factors Influencing Employee Attrition and Develop Insights for Targeted Retention Strategies.

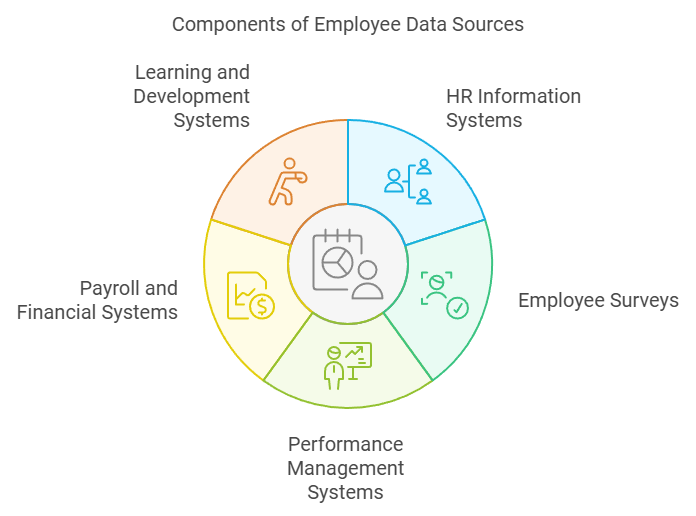
**Data Requirement:**

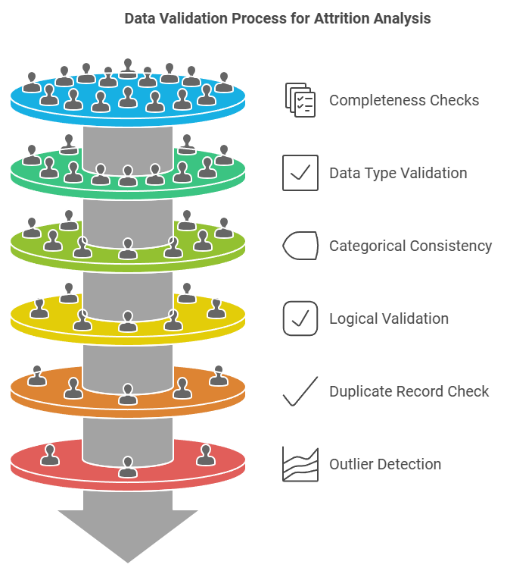
To address this problem statement effectively, the analysis would require the following key data components from the dataset:



**Data Collection:**

1. **HR Information Systems (HRIS)**: The primary source for employee demographics, compensation, job roles, department information, and employment history. HRIS platforms like Workday, SAP SuccessFactors, or Oracle PeopleSoft store extensive employee data that is vital for attrition analysis.
2. **Employee Surveys**: **Job Satisfaction** and **Environment Satisfaction** metrics are often collected through periodic employee engagement surveys. These surveys may also capture data on work-life balance, relationship satisfaction, and other subjective factors influencing employee sentiment.
3. **Performance Management Systems**: Systems that track **performance ratings** and **promotion history** provide information on how employees are evaluated and their career progression. This data can help determine if performance-related aspects contribute to attrition.
4. **Payroll and Financial Systems**: Payroll data provides information on **monthly income**, **salary hikes**, **stock options**, and **overtime** hours. Financial systems that store this data are often integrated with HR systems to keep track of compensation and benefits.
5. **Learning and Development Systems**: To assess **training times** and development opportunities, data from Learning Management Systems (LMS) can be leveraged. This data shows how often employees participate in training and development, potentially influencing retention.



**Data Validation:**

To ensure the data’s accuracy, consistency, and completeness for attrition analysis, data validation is essential. Here’s a structured approach for validating key data points in this type of dataset:

1. Completeness Checks

- Missing Values: Check each column for missing values, especially in critical fields such as "Attrition," "Job Role," "Monthly Income," "Performance Rating," and "Satisfaction Metrics."

- Field Dependency: Ensure that related fields contain data (e.g., if "Years at Company" is greater than zero, fields like "Job Role" and "Department" should not be blank).

- Employee Count Consistency: Verify that the number of records aligns with the actual employee count to ensure no records are missing.

2. Data Type Validation

- Consistent Data Types: Check that all columns have appropriate data types (e.g., "Age" and "Monthly Income" should be numeric, while "Attrition" and "Department" should be categorical).

- Range Validation: Ensure that numeric fields fall within realistic ranges. For example:

- Age: Within a typical working range (18-65).

- Monthly Income: Falls within reasonable bounds for the company.

- Years at Company: Is non-negative and doesn’t exceed an improbable number (e.g., < 40 years).

3. Categorical Consistency

- Standardized Values: Ensure categorical columns like "Department," "Job Role," and "Business Travel" use consistent naming conventions and spellings. For example, variations like "Sales" vs. "sales" or "Travel\_Rarely" vs. "Rarely Travels" should be standardized.

- Expected Categories: Verify that each categorical field includes only expected values. For example, "Performance Rating" might have defined levels (1-5), so any value outside this range would need investigation.

4. Logical Validation

- Business Logic Checks:

- Attrition Consistency: If an employee's "Attrition" status is "Yes," fields like "Years at Company" or "Years in Current Role" should not be updated beyond their exit date.

- Promotion Validity: For "Years Since Last Promotion," ensure that values are less than or equal to "Years at Company."

- Total Working Years: Should always be greater than or equal to "Years at Company."

- Relationship Consistency: Ensure values like "Total Working Years" match logical assumptions, such as being greater than or equal to "Years in Current Role" and "Years with Current Manager."

5. Duplicate Record Check

- Unique Identifiers: Verify that each employee has a unique identifier (e.g., "Employee Number") and check for duplicates in this field. Any duplicated records would indicate a data entry issue.

6. Outlier Detection

- Extreme Values: Use statistical methods (e.g., Z-score or IQR) to detect extreme outliers in numeric fields like "Monthly Income" and "Years at Company" that may indicate data entry errors.

- Logical Outliers: For example, an employee with "Years in Current Role" greater than "Years at Company" would be an obvious error.

7. Cross-System Verification

- Audit Against Source Data: Where possible, cross-check against the original source systems (e.g., HRIS, Payroll) to ensure that data matches for sensitive fields like "Monthly Income" and "Performance Rating."

- Survey Data Cross-Validation: If satisfaction data is from surveys, confirm that survey responses align with HR records for each employee ID.

8. Data Imputation and Cleaning

- After validation, handle missing or incorrect values by:

- Imputing Missing Data: Where feasible, using mean, median, or mode imputation for numeric or categorical fields.

- Correction of Errors: For identified inconsistencies, correct values based on reasonable assumptions or historical data.

**Data Preprocessing:**

1. **Data Cleaning:** Handle missing values through imputation or removal, ensure categorical consistency, and address outliers.
2. **Feature Engineering**: Convert "Attrition" to binary, create tenure and satisfaction composite features, and calculate work-life balance scores numerically.
3. **Encoding Categorical Variables:** Label encode binary variables and one-hot encode multi-class fields like "Job Role."
4. **Feature Scaling:** Normalize or standardize numerical features like "Monthly Income" and "Years at Company."
5. **Handle Class Imbalance:** Balance the dataset using techniques like SMOTE or under-sampling the majority class.
6. **Transformation**: Apply log transformation to skewed variables and bin continuous features if useful for interpretation.
7. **Feature Selection:** Remove redundant features based on correlation and rank feature importance to focus on predictive variables.
8. **Data Splitting**: For modeling, use a train-test split and cross-validation for accurate model evaluation.

**Tools:**

**1)** **Power BI**:

* A Microsoft tool for creating interactive data visualizations and dashboards.
* Helps in visualizing trends, patterns, and correlations within the data, making it easier to identify issues (e.g., class imbalance) and share insights with stakeholders.

**2)** **Tableau**:

* A powerful data visualization tool used for creating detailed visual analytics.
* Allows you to build quick, interactive visualizations that reveal insights, such as distribution patterns and outliers, critical for exploratory data analysis (EDA).

**3)** **Advanced Excel**:

* Excel skills beyond the basics, including data cleaning, pivot tables, formulas, and functions like VLOOKUP and macros.
* Useful for data cleaning, transformation, and initial analysis. Excel is great for handling missing values, creating summary tables, and performing quick data checks.

**4)** **Python**:

* A versatile programming language often used in data science, especially with libraries like Pandas, NumPy, and Scikit-learn.
* Essential for automating data cleaning, performing advanced preprocessing steps (e.g., feature engineering, scaling), and building predictive models. Python’s libraries make complex preprocessing and analysis easier and repeatable.

**5)** **SQL**:

* Structured Query Language, used to manage and query relational databases.
* Ideal for extracting, filtering, and merging data from databases, ensuring data consistency before analysis. SQL is particularly useful in handling large datasets stored in databases, allowing you to select only relevant data for analysis.

**Dashboard**

A dashboard is an essential tool for summarizing and visualizing data, making complex information accessible and actionable. In an employee attrition analysis, a dashboard allows HR and management to monitor critical metrics, like attrition rates, job satisfaction, and tenure trends, in real time. By presenting data through visual elements—such as graphs, charts, and tables—a dashboard reveals patterns and trends that may not be immediately apparent in raw data, such as attrition spikes in specific departments or age groups. This enables quick, data-driven decision-making and allows stakeholders to drill down into specific areas, like high-risk employee segments or satisfaction scores, to investigate potential causes. Additionally, dashboards make it easier to communicate insights to a broader audience, helping others understand and engage with findings for better-informed strategic actions.