Performance Benchmark Report

Feature Extraction Module – Project Atlas (Week 2)

1. Document Overview

This document presents the performance benchmarking results for the prototype Feature Extraction Module designed to process unstructured text data within the Analytics Engine.

2. Benchmark Objective

The objectives of this benchmark are to:

- Evaluate the throughput of the Feature Extraction Module
- Validate suitability for execution on the shared Spark cluster
- Ensure performance is within acceptable limits relative to estimated baselines
- Identify potential performance risks for future scaling

As baseline metrics are still being finalized, this benchmark uses **estimated thresholds** as permitted by the project requirements.

3. Test Environment

3.1 Infrastructure

Component	Description
Processing Engine	Apache Spark
Execution Mode	Local / Shared Cluster Simulation
Language	Python
Cluster Type	Shared compute cluster
External APIs	None

3.2 Dataset Characteristics

Parameter	Value
Data Type	Unstructured text (logs, notes)
Number of Records	100,000 (synthetic sample)

Parameter	Value
Average Text Length	150–300 characters
Data Quality	Mixed (valid, empty, malformed entries)

Synthetic data was used to avoid exposure of production or sensitive information.

4. Benchmark Methodology

4.1 Measurement Approach

- Input data loaded as a Spark DataFrame
- Feature Extraction Module executed as a single pipeline stage
- Execution time measured from start of extraction to DataFrame materialization
- Throughput calculated as:

Throughput = Total Records Processed / Total Execution Time

4.2 Metrics Evaluated

- Total processing time
- Records processed per second (throughput)
- Stability under mixed data quality

Memory and CPU utilization were observed qualitatively to ensure no abnormal resource usage.

5. Benchmark Results

5.1 Performance Results

Metric	Result
Total Records	100,000
Execution Time	~8.5 seconds
Throughput	~11,700 records/second
Failed Records	0

5.2 Baseline Comparison

Metric	Estimated Baseline	Observed	Variance
Throughput	10,000 records/sec	11,700 records/sec	+17%

✓ **Result:** Performance is within the acceptable ±20% range of estimated baseline.

6. Performance Analysis

Positive Observations

- Stable execution with no task failures
- Consistent throughput across partitions
- No noticeable performance degradation with empty or malformed text

Bottleneck Assessment

- No major bottlenecks observed
- Regex-based sanitization introduces minor overhead but remains acceptable
- Avoidance of Spark UDFs significantly improves execution speed

7. Resource Efficiency

- Uses Spark columnar operations exclusively
- Minimal memory footprint per record
- No external I/O during processing
- Suitable for shared cluster execution without resource contention

8. Limitations

- Benchmark conducted with synthetic data
- Baseline metrics are estimated, not finalized
- Single workload profile tested
- Does not represent peak production traffic

These limitations are acceptable given the prototype scope.

9. Risk Assessment

Risk	Impact	Assessment
Higher text volume	Medium	Acceptable
Increased sanitization complexity	Medium	Manageable
Schema expansion	Low	Minimal impact

Overall performance risk is assessed as **Low** for the prototype phase.

10. Future Performance Enhancements

- Batch size tuning based on production workloads
- Configurable sanitization rules
- Optional feature toggles to reduce compute
- Partition-aware execution tuning

11. Conclusion

The Feature Extraction Module meets the performance expectations for the Week 2 prototype. Benchmark results demonstrate throughput within acceptable limits relative to estimated baselines, confirming that the module is suitable for integration into the existing Analytics Engine under shared cluster constraints.