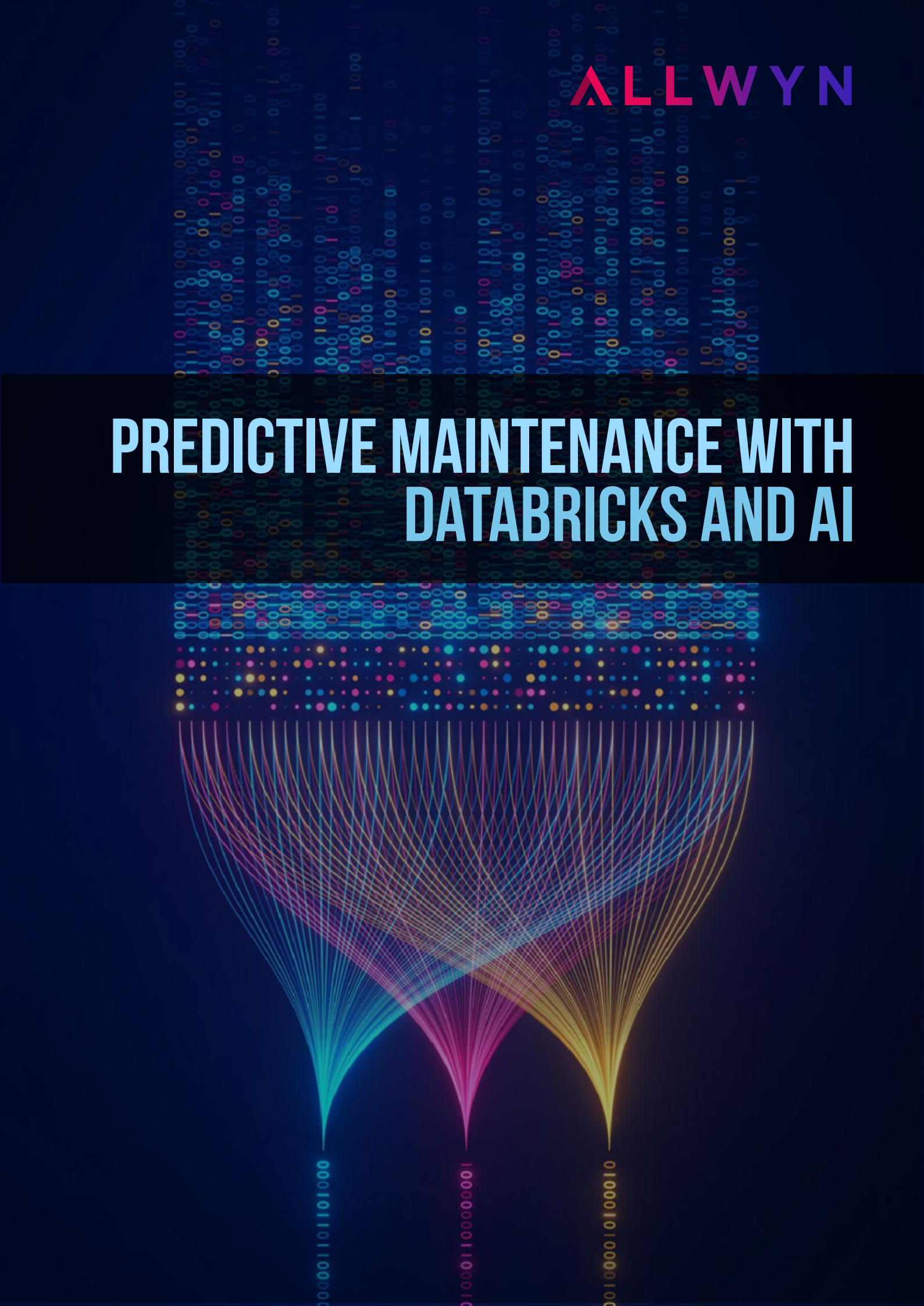
ALLWYN

PREDICTIVE MAINTENANCE WITH DATABRICKS AND AI



Summary

Railroad organizations today face increasing pressure to improve on-time performance, reduce equipment downtime, and optimize maintenance operations. Traditional approaches rely on reactive or scheduled maintenance, which often leads to unnecessary costs, service interruptions, and inefficient use of resources.

This whitepaper presents a Proof of Concept (POC) conducted by Allwyn using the Databricks Lakehouse Platform integrated with AWS S3. The POC showcases how machine learning (ML) and artificial intelligence (AI) can enable predictive maintenance across the railroad's equipment fleet. By ingesting and analyzing telemetry, maintenance, and error datasets, the team demonstrated accurate availability calculations, proactive maintenance insights, and predictive model ing for equipment health.

The results illustrate how predictive analytics can transform railroad maintenance practices by:

- **Increasing equipment** availability and reliability.
- **Reducing maintenance costs** through early detection.
- **Providing real-time dashboards** and automated alerts for engineering teams.
- **Enhancing passenger** confidence via improved on-time performance.

Introduction

Why Predictive Maintenance Matters?

Rail transportation is a complex system where mechanical reliability directly impacts passenger satisfaction, safety, and revenue. Equipment delays and failures affect schedules across the nationwide routes, costing millions annually in unplanned downtime and repairs.

Traditional maintenance falls into two categories:

Reactive Maintenance:

Addressing issues only after they occur, leading to costly delays and secondary damage.

Preventive Maintenance:

Addressing issues only after they occur, leading to costly delays and secondary damage.

Predictive Maintenance (PdM), powered by real-time data and ML models, bridges these gaps. By monitoring telemetry streams and analyzing historical failure patterns, predictive systems forecast when a component is likely to fail—optimizing interventions before costly bre cur. The adoption of predictive maintenance helps transit organizations to enhance operational efficiency, safety, and customer experience

POC Objectives:

01 **Demonstrate Databricks Capabilities –** Prove the Lakehouse platform can integrate diverse datasets (telemetry, maintenance, errors) and deliver predictive analytics at scale.

02 **Showcase Business Value –** Highlight how predictive maintenance reduces downtime, optimizes costs, and improves service reliability.

03 **Provide a Roadmap for Adoption –** Establish a phased approach for the national railroad to operationalize predictive analytics within

Data Sources & Environment

Datasets

The POC leveraged representative public datasets reflecting railroad operations:

- **Telemetry Data:** Hourly averages of voltage, vibration, rotation, and pressure for 100 machines over 1 year.
- **Machines Data:** Machine ID, model, and age distribution (0–20 years).
- **Error Data:** 3,919 records of non-fatal errors across 4 error types.
- **Maintenance Records:** 3,286 entries detailing proactive and reactive component replacements.
- **Failure Data:** 761 failure-driven replacements (subset of maintenance records).

Technical Architecture

The POC leveraged representative public datasets reflecting railroad operations:

- **Data Storage:** AWS S3 bucket for raw datasets.
- **Processing Platform:** Databricks Lakehouse (PySpark for ETL, SQL for queries).
- **Integration:** Databricks cluster mounted to S3 for direct access.
- **Visualization:** Tableau dashboards for KPIs and insights.
- **ML Deployment:** Databricks MLflow for model versioning, serving, and REST API deployment.

POC Implementation Process

1. Data Ingestion & Preprocessing

- Mounted S3 storage into Databricks.
- Read CSV datasets into Spark DataFrames.
- Created Hive tables (allwynpocdb) for structured querying.
- Conducted data cleaning, duplicate checks, and normalization.
- Feature engineering: rolling averages, lagged features, and failure

2. Exploratory Data Analysis (EDA)

EDA provided critical insights into machine behavior:

- No clear correlation between machine age and failures—dispelling assumptions that older machines always fail more frequently.
- Errors did not always lead to failures, indicating nuanced relationships between telemetry and breakdown events.
- Maintenance records revealed the split between proactive (scheduled) vs reactive (failure-driven) interventions.
- Higher maintenance activity was observed in 2015 compared to 2014, suggesting evolving operational conditions.

3. Availability Calculation

The team built an aggregate availability table with the following schema:

MachineID | Telemetry_Date | Total_Time | Healthy_Time | Error_Time | Availability (%)

Formula:

Availability : Healthy_time - Error_time)/(Total_time)

Performance : Healthy_time/ Total_time

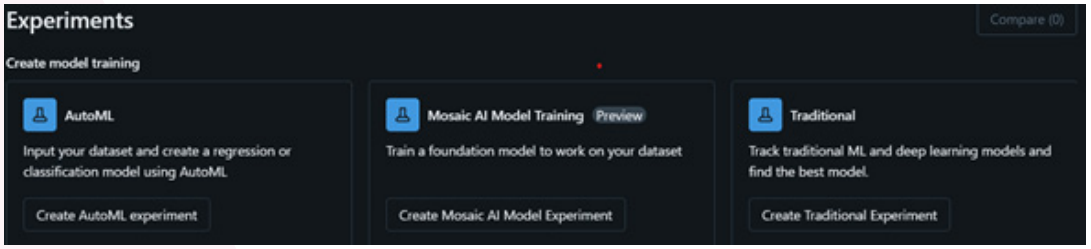
Quality: (Total Parts Made - Defective Parts Made)/Total Parts Made

This enabled daily availability tracking for each of the 100 machines, forming the baseline KPI for predictive maintenance.

4. Machine Learning & Forecasting

The team built an aggregate availability table with the following schema:

- **Approach:** Time-series forecasting with Databricks AutoML

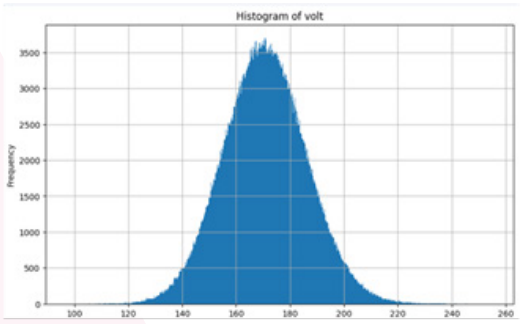


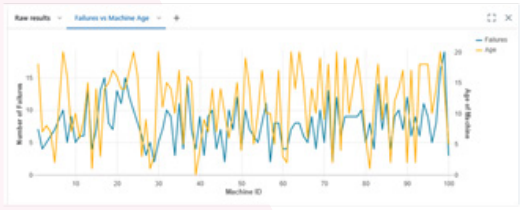
- **Algorithms:** ARIMA, Prophet (tested for accuracy using SMAPE, RMSE, and MSE).
- **Objective:** Forecast machine availability and detect patterns leading to stoppages.
- **Deployment:** Best-performing model deployed via MLflow, accessible through REST APIs.

5. Visualization & Reporting

- Tableau dashboards displayed KPIs such as availability trends, failure predictions, and maintenance workloads.
- Automated reporting pipelines ensured engineering teams had real-time visibility.

Key Findings





1. Equipment Behavior: Age alone is not a predictor of failures—data-driven insights are essential.
2. Error Patterns: Errors occur frequently but don't necessarily indicate imminent breakdowns, reinforcing the need for predictive modeling.
3. Maintenance Insights: Distinction between proactive vs reactive events clarifies cost-saving opportunities for scheduled interventions.
4. Data Reliability: Telemetry captured consistent, non-duplicated, and normally distributed values across attributes (voltage, vibration, rotation, pressure).

Results & Benefits

Technical Architecture

- **Real-Time Monitoring:** Engineering teams gained visibility into machine health at daily granularity.
- **Predictive Scheduling:** Maintenance could be forecasted and planned, reducing unexpected breakdowns.
- **Improved On-Time Performance (OTP):** Enhanced train reliability translates into higher rider confidence and reduced service delays.

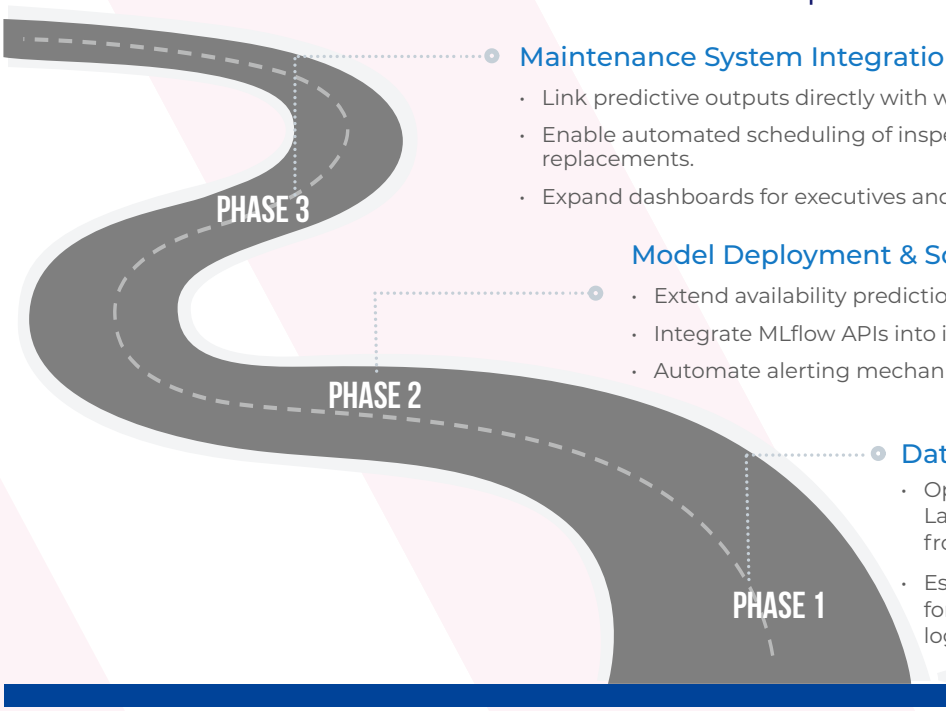
Financial Benefits

- **Cost Reduction:** Optimized maintenance schedules lower labor, parts, and downtime costs.
- **Resource Allocation:** Engineering teams can prioritize interventions based on predictive insights.
- **ROI Potential:** The POC indicates significant savings if scaled fleet-wide.

Strategic Benefits

- **Modernized Analytics:** Transition from static data warehouses to a self-service Lakehouse model.
- **Scalable Architecture:** Designed for future expansion with real-time IoT sensor integration.
- **AI Foundation:** Establishes groundwork for broader AI initiatives (e.g., demand forecasting, energy optimization).

Recommendations & Roadmap



Maintenance System Integration

- Link predictive outputs directly with work order management.
- Enable automated scheduling of inspections and component replacements.
- Expand dashboards for executives and operations leadership.

Model Deployment & Scaling

- Extend availability prediction models across the full fleet.
- Integrate MLflow APIs into internal maintenance systems.
- Automate alerting mechanisms for critical failures.

Data Infrastructure Setup

- Operationalize the Databricks Lakehouse with real-time pipelines from onboard train sensors.
- Establish data governance policies for telemetry and maintenance logs.

Success Metrics

- Reduction in downtime hours.
- Lower maintenance costs (proactive vs reactive ratio).
- Improved on-time performance benchmarks.

Future Opportunities

1. **IoT Integration:** Direct ingestion of sensor streams from locomotives into Databricks.
2. **Generative AI for Maintenance Logs:** Use Mosaic AI (Databricks' GenAI platform) to interpret technician notes and identify root causes.
3. **Cross-Agency Benchmarking:** Share PdM insights across federal transportation entities for system-wide reliability improvements.
4. **Advanced Modeling:** Incorporate deep learning approaches (e.g., LSTMs, transformers) for complex sequence predictions.

Conclusion

The POC validated that predictive maintenance powered by **Databricks AI/ML** is both technically feasible and operationally valuable for any railroad organization. By calculating daily equipment availability, forecasting failures, and delivering actionable dashboards, the initiative showcased significant potential for **cost reduction, operational efficiency, and customer experience improvement**.

Scaling this solution across the fleet will modernize maintenance operations, drive measurable ROI, and set a foundation for broader AI-driven transformation in U.S. rail systems.

Contact Information

Sid K. Hasan
Chief Growth Officer (CGO),
Allwyn Corporation
+1 (415) 377-0693
sid.hasan@allwyncorp.com

Swathi Young
Chief Technology Officer (CTO),
Allwyn Corporation
+1 (703) 638-2538
swathi.young@allwyncorp.com

About Allwyn Corporation

Allwyn Corporation (www.allwyncorp.com) is a forward thinking, innovative software solutions company, headquartered in the Metropolitan Washington DC area. Allwyn was founded in 2003 with a mission to help organizations address complex technology challenges by providing industry-leading tools, technologies, seasoned professionals, and proven methodologies. We are proud to be certified for ISO 9001 (Quality), ISO 27001 (Security), and ISO 20000 (Service Delivery).
With a team of ~200 professionals, Allwyn delivers high-quality services to a wide range of clients in the public and private sector.

Allwyn has been providing leading-edge IT professional services to various government agencies through the GSA MAS Schedule. We are also on the FAA eFAST, GSA OASIS+, and GSA STARS III contract vehicles.

Allwyn has experience with implementing Artificial Intelligence and Machine Learning solutions and Modernizing Applications using Low Code Technologies. Our relationships with AWS, Appian, ServiceNow, Microsoft, Databricks, Informatica, Salesforce, etc. strengthen our ability to support our customers in their Digital Transformation journey. We are already supporting several of our customers in the public as well as commercial sector with their cloud adoption strategies and Artificial Intelligence and Machine Learning implementations. For additional information on Allwyn's full range of services, please visit our website at www.allwyncorp.com.