DIPLOM/MASTERARBEITSTHEMA:
ASFINAG DATA LAKE

Der weltweite Datenbestand wächst jährlich um ungefähr 40 %, was einer Verdoppelung der Daten zirka alle zwei Jahre entspricht. Laut Prognose von International Data Corporation (IDC) wird das Datenvolumen auf der Erde bis zum Jahre 2020 auf 44 Zettabyte (ZB) ansteigen.

Dieser Datenzuwachs, auch wenn er in einem kleineren Kontext anzutreffen ist, spielt bei der ASFINAG eine Rolle. Obwohl die Datenmengen für die ASFINAG überschaubar sind, ist die Entwicklung kostengünstiger Architekturen für die Zukunft notwendig, sowohl um Daten persistent abzuspeichern als auch um Datenanalysen im großen Stil betreiben zu können. Zurzeit bereiten Datensilos durch die isolierte Datenhaltung enorme Hindernisse bei der Datenhaltung, Datenverarbeitung sowie Datenanalyse. Die redundante Datenhaltung, die teilweise in verschiedenen Informationssystemen notwendig ist, führt auch dazu, dass das Konzept des Single Point of Truth (SPOT) im Unternehmen verloren geht.

Ziel der Diplomarbeit

Im Zuge der Masterarbeit / Diplomarbeit soll neben dem Aufbrechen der Datensilos ein Architekturvorschlag für ein ASFINAG - Data Lake geliefert werden. Dabei sollte der Fokus auf Hadoop und das Ökosystem liegen.

Die Masterarbeit / Diplomarbeit soll des Weiteren für folgende konkrete Fragestellungen eine Basis liefern:

- Analyse, welche weiterverbreiteten Komponenten im Hadoop-Ökosystem in Zusammenspiel mit aktuell im Produktiveinsatz befindlichen Systemen der ASFINAG geeignet sind. Daraus abgeleitet soll ein Konzept für ein Data Lake erarbeitet werden, das die Anforderungen der ASFINAG in Hinblick auf eine Analyse- und Simulationsplattform erfüllt.
- Das Aufbrechen der Datensilos und die Errichtung eines Data Lake rollt das Problem des Data Governance neu auf. Dazu müssen Strategien und/oder Komponenten im Hadoop Umfeld evaluiert werden. Vor allem ist das Erkennen von zusammenhängenden Daten sehr wichtig. Diesbezüglich soll mithilfe der Data Governance - Strategie das Auffinden, Löschen und Anonymisieren von zusammenhängenden Daten im Data Lake ermöglicht bzw. vereinfacht werden.