










DATA SCIENCE PIPELINE (TOOLS AND TECHNOLOGIES)

	PROCESS	STEPS NEEDED	TOOLS & LANGUAGES	ROLE
	Collection / Data Stream	<ul style="list-style-type: none"> Scrapers Internal Data Collection Other Databases Survey 	<ul style="list-style-type: none"> Amazon Kinesis. Google Cloud DataFlow. Azure Stream Analytics. IBM Streaming Analytics. Apache Storm. Striim. StreamSQL. Redis Apache Spark 	<ul style="list-style-type: none"> Data Engineer Data Analyst Business Analyst
	Storing Data	<ul style="list-style-type: none"> Access Management / Governance Security Database Creation Entity Relationships 	<ul style="list-style-type: none"> Oracle Database Amazon RDS Google Cloud Bucket MongoDB MySQL PostgreSQL SQL Server Teradata Hadoop Cluster 	<ul style="list-style-type: none"> Data Engineer
	Retrieving Data	<ul style="list-style-type: none"> Extract Analytics on Query CSVs / local storage 	<ul style="list-style-type: none"> SQL Hive Google Big Query 	<ul style="list-style-type: none"> Data Engineer Data Analyst Business Analyst Data Scientist
	Cleaning	<ul style="list-style-type: none"> Null values Renaming String Manipulation Calculating new variables from existing Data Types Drop columns / Rows Distributions Duplicates Normalization / Standardization Data Imputation Unit Tests 	<ul style="list-style-type: none"> Python R SQL Alteryx SAS Oracle ADS Amazon Glue 	<ul style="list-style-type: none"> Data Analyst Business Analyst Data Scientist
	Joining	<ul style="list-style-type: none"> Merge Tables Post Merge Cleaning 	<ul style="list-style-type: none"> Python R SQL Alteryx Amazon Glue 	<ul style="list-style-type: none"> Data Analyst Business Analyst

DATA SCIENCE PIPELINE (TOOLS AND TECHNOLOGIES)

		<ul style="list-style-type: none"> • SAS • Oracle ADS 	<ul style="list-style-type: none"> • Data Scientist
	Exploratory Analysis <ul style="list-style-type: none"> • Visualization • Basic Statistics • Population Distributions • Correlations • Feature Selection 	<ul style="list-style-type: none"> • Python • R • Tableau • Power BI • Qrvey • Splunk • SPSS • Oracle ADS • WEKA 	<ul style="list-style-type: none"> • Data Analyst • Business Analyst • Data Scientist
	Modeling <ul style="list-style-type: none"> • Modeling Technique / Assumptions • Test Design • Parameter Tuning • Model Creation • Model Assessment / Testing 	<ul style="list-style-type: none"> • Python • R • Oracle ADS • Amazon Sage Maker • Alteryx • Tensorflow • Pytorch • Keras • Sklearn • MATLAB • Shogun • Apache Spark MLLib • Apache Mahout • WEKA • Google Cloud ML • H2O.ai 	<ul style="list-style-type: none"> • Data Scientist • Machine Learning Engineer
	Deployment <ul style="list-style-type: none"> • Monitoring • Docker Images • Cloud • Rest API / API • Logging 	<ul style="list-style-type: none"> • Oracle Cloud • Google Cloud • Amazon Web Services • Azure • Docker • Heroku • Pickle 	<ul style="list-style-type: none"> • Data Scientist • Machine Learning Engineer
	Visualization <ul style="list-style-type: none"> • Dashboarding • Reporting • Website Integration 	<ul style="list-style-type: none"> • Python (Django / Flask) • R (Shiny) • Tableau • Power BI • Qrvey • Splunk • Sisense • Qlikview • Oracle Business Intelligence • Oracle Analytics Cloud 	<ul style="list-style-type: none"> • Data Scientist • Business Analyst • Data Analyst