电网数据中台开源技术栈清单

以下表格列出了数据中台各层次可采用的开源软件技术栈，按功能模块进行组织：

层级	功能模块	推荐开源软件	开源替代方案	说明
1. 数据源接入层	实时数据采集	Apache Kafka	RabbitMQ, Pulsar	高吞吐消息队列，适合实时数据采集
	协议适配	NiFi	StreamSets, Logstash	支持多种工业协议转换
	批量数据导入	Apache Airflow	Luigi, Prefect	工作流编排与调度
	元数据管理	Apache Atlas	DataHub, Amundsen	元数据目录与血缘管理
	API管理	Kong	Apigee Open Source, Tyk	API网关与管理
2. 数据处理层	数据清洗	Apache Spark	Flink, Beam	分布式数据处理框架
	数据转换	dbt (data build tool)	Apache Airflow, Talend Open Studio	SQL优先的数据转换工具
	数据质量	Great Expectations	Deequ, Griffin	数据质量验证框架
	数据血缘	OpenLineage	Marquez, DataHub	数据血缘追踪
	数据治理	Apache Ranger	Sentry, OPA	访问控制与权限管理
	规则引擎	Drools	Apache Camel, Easy Rules	业务规则执行引擎
3. 数据存储与分析层	数据湖	Apache Hadoop	MinIO, Ceph	分布式存储系统
	数据湖格式	Delta Lake	Apache Hudi, Apache Iceberg	事务性数据湖格式
	数据仓库	Greenplum	Apache Doris, ClickHouse	MPP数据仓库
	时序数据库	TimescaleDB	InfluxDB, OpenTSDB	电力时序数据存储
	OLAP引擎	Apache Kylin	Apache Druid, Apache Pinot	OLAP多维分析
	SQL查询引擎	Presto/Trino	Apache Spark SQL, Apache Impala	跨源SQL查询
	数据可视化	Apache Superset	Redash, Grafana	自助式BI与数据探索
	报表工具	Metabase	JasperReports, BIRT	报表设计与生成
4. AI模型准备层	特征工程	Feast	MLflow, Hopsworks	特征存储和管理
	向量数据库	Milvus	Qdrant, Weaviate	高维向量存储与检索
	数据标注	Label Studio	CVAT, Doccano	多模态数据标注平台
	数据增强	Albumentations	imgaug, TorchIO	数据增强库
	样本管理	DVC (Data Version Control)	LakeFS, Pachyderm	数据版本控制
5. 模型开发与部署层	机器学习框架	Scikit-learn	XGBoost, LightGBM	传统机器学习算法
	深度学习框架	PyTorch	TensorFlow, MXNet	深度学习框架
	模型训练平台	MLflow	Kubeflow, H2O	模型生命周期管理
	超参数优化	Optuna	Ray Tune, Hyperopt	超参数自动优化

层级	功能模块	推荐开源软件	开源替代方案	说明
	模型仓库	MLflow Model Registry	Seldon Core, BentoML	模型版本管理与服务化
	模型解释性	SHAP	LIME, ELI5	模型解释工具
	模型部署	Seldon Core	KServe, BentoML	Kubernetes模型服务部署
	模型监控	Prometheus	Grafana, Evidently	模型性能监控
	API服务	FastAPI	Flask, Streamlit	轻量级API开发框架
通用基础设施	容器编排	Kubernetes	Docker Swarm, Nomad	容器管理与编排
	服务网格	Istio	Linkerd, Consul	微服务通信管理
	任务调度	Apache Airflow	Argo Workflows, Jenkins	工作流编排与调度
	分布式文件系统	HDFS	MinIO, Ceph	大规模数据存储
	身份认证	Keycloak	OAuth2 Proxy, Dex	身份验证与SSO
	监控告警	Prometheus + Grafana	Zabbix, Nagios	系统监控与可视化
	日志管理	ELK Stack	Graylog, Loki	日志收集与分析
	代码版本控制	GitLab	GitHub, Gitea	源代码管理
	CI/CD	GitLab CI	Jenkins, ArgoCD	持续集成与部署

补充说明：

1. **开源优先**：所有推荐的软件均为开源产品，可根据南方电网实际情况进行选型
2. **技术集成**：部分功能可能需要多个工具组合使用以实现完整能力
3. **电力行业适应性**：选型时考虑了电力行业高可靠性、大数据量的特点
4. **技术成熟度**：优先选择社区活跃、技术成熟的开源项目
5. **替代方案**：提供了备选方案，可根据团队技术栈和经验灵活调整

此技术栈可作为项目初期技术选型的基础，后续可根据具体需求和性能测试结果进行调整优化。