Microsoft Azure-ի և ամպային տեխնոլոգիաներին վերաբերվող հիմնական հասկացությունները

Ներածություն

Ամպային տեխնոլոգիաները հանդիսանում են ժամանակակից ծառայությունների մատուցման մոդել, որի հաշվարկային և ծրագրային ռեսուրսները հասանելի են համացանցի միջոցով՝ ըստ պահանջի: Այս մոտեցումը ազատում է կազմակերպություններին սեփական սերվերների, տվյալների պահպանման համակարգերի և ցանցային ենթակառուցվածքի ստեղծման ու սպասարկման դժվարություններից: Ամպային ծառայությունները ներառում են վիրտուալ սերվերներ և հաշվարկային հզորություն, տվյալների պահեստավորում և արխիվացում, տվյալների բազաներ ու վերլուծական ծառայություններ, ցանցային գործիքակազմ՝ firewall, VPN, ինչպես նաև ծրագրային ապահովում՝ DevOps միջավայրեր: Ամպային հաշվարկի հիմնական իմաստը կայանում է նրանում, որ բոլոր ռեսուրսները հասանելի են ցանկացած տեղ՝ համացանցի առկայության դեպքում, և վճարումն իրականացվում է օգտագործված ռեսուրսների չափով:

Ամպային մոդելի կարևորագույն առավելություններից է մասշտաբելիությունը, որը թույլ է տալիս արագ ավելացնել կամ նվազեցնել օգտագործվող ռեսուրսների քանակը՝ ըստ անհրաժեշտության, ինչը հատկապես կարևոր է սեզոնային կամ անկանոն բեռնվածություն ունեցող բիզնեսների համար: Մյուս կարևոր առավելությունն է ցածր նախնական ծախսերը. ընկերությունները չեն գնում թանկարժեք սարքավորումներ, չեն կատարում սպասարկման վճարներ և նվազեցնում են IT անձնակազմի ծանրաբեռնվածությունը, քանի որ համակարգը գործում է «վճարիր օգտագործելուց հետո» սկզբունքով: Բացի այդ, ամպային համակարգերը ապահովում են բարձր հուսալիություն, քանի որ տվյալների կենտրոնները, որոնք բաշխված են աշխարհով մեկ, ապահովում են ծառայությունների անընդհատ աշխատանքը, և եթե մի կենտրոնում խնդիր առաջանա, բեռնվածությունն ավտոմատ կերպով տեղափոխվում է մյուս կենտրոնների վրա:

Անվտանգության առումով ամպային հարթակները ներառում են ներկառուցված մեխանիզմներ՝ մշտական մոնիթորինգ, տվյալների կոդավորում, վերահսկում և պաշտպանություն DDoS հարձակումներից, որոնք համապատասխանում են միջազգային ստանդարտներին, ինչպիսիք են ISO, GDPR և HIPAA: Ամպային ծառայությունները նաև ապահովում են արագ զարգացում և ավտոմատացում՝ տրամադրելով CI/CD գործընթացներ, ավտոմատացման գործիքներ և թեստավորման համակարգեր, ինչը արագացնում է ծրագրային ապահովման թողարկումը:

Microsoft Azure

Microsoft Azure-ը Microsoft ընկերության ստեղծած լայնածավալ և բազմաֆունկցիոնալ ամպային հարթակն է, որը առաջարկում է ավելի քան 200 ծառայություն և համապատասխանում է ինչպես փոքր ընկերությունների, այնպես էլ մեծ ձեռնարկությունների պահանջներին: Azure-ը հնարավորություն է տալիս կառուցել, փորձարկել, տեղակայել և կառավարել ծրագրեր՝ առանց ֆիզիկական ենթակառուցվածքներ ունենալու անհրաժեշտության:

Azure-ի հիմնական ծառայությունները ներառում են հաշվողական հզորություն, ինչպիսիք են Virtual Machines (VMs), որոնք տրամադրում են տարբեր օպերացիոն համակարգերով վիրտուալ սերվերներ, Azure Functions՝ առանձ սերվերների հաշվարկ, որտեղ կոդը աշխատում է իրադարձությունների հիման վրա, և App Services՝ վեբ ծրագրերի հոստինգի համար: Տվյալների պահեստավորման ոլորտում Azure-ն առաջարկում է Blob Storage, որը նախատեսված է մեծ չստրուկտուրավորված տվյալների պահպանման համար, Azure Files՝ կիսվող ֆայլային համակարգ, ինչպես նաև Data Lake Storage՝ մեծ չափի պահեստ տարբեր տիպի տվյալների համար:

Տվյալների բազաների ոլորտում ծառայությունների շարքում են Azure SQL Database, որը հանդիսանում է կառավարվող տվյալների բազա, Cosmos DB՝ գլոբալ բաշխված և բարձր աշխատանքային կարողություններով NoSQL տիպի տվյալների բազա, և PostgreSQL/MySQL ծառայություններ: Մեքենայական ուսուցման և արհեստական բանականության համար Azure-ը տրամադրում է Azure Machine Learning ծառայությունը՝ մեքենայական ուսուցման մոդելների մշակման, ուսուցման և տեղակայման ամբողջական միջավայր, ինչպես նաև Cognitive Services՝ տեսողության, խոսքի, լեզվի, որոնման և այլ տիպերի API-ների համախումբ:

Տվյալների պրոցեսավորման և ինտեգրման ոլորտում Azure-ի ծառայությունները ներառում են Azure Data Factory-ն՝ տվյալների տեղափոխման և փոխակերպման համար, և Azure Synapse Analytics՝ մեծածավալ տվյալների վերլուծության և բիզնես ինտելեկտի ապահովման համար: Ցանցային ծառայությունների շարքում ընդգրկված են Virtual Network, Load Balancer, VPN Gateway, Firewall, ինչպես նաև CDN, DDoS հարձակումներից պաշտպանություն և այլ ցանցային գործիքներ:

Azure-ի հիմնական առավելությունն այն է, որ այն ապահովում է ճկունություն, գլոբալ հասանելիություն և անվտանգություն, միաժամանակ նվազեցնելով կազմակերպությունների տեխնիկական և ֆինանսական բեռնվածությունը:

ETL/ELT տվյալների ինտեգրման եղանակներ

ETL-ը տվյալների ինտեգրման դասական գործընթաց է, որի նպատակը տարբեր աղբյուրներից ստացված տվյալները մաքրելն, փոխակերպելն ու պահպանելն է տվյալների պահեստում: ETL-ի գործընթացը բաղկացած է երեք փուլից՝Extract, Transform և Load:

Առաջին փուլը՝ Extract, որի ընթացքում տվյալները վերցվում են տարբեր աղբյուրներից, օրինակ՝ տվյալների բազաներից, ինչպիսիք են SQL Server կամ Oracle, API-ներից և վեբ ծառայություններից, CSV, JSON կամ Excel ֆայլերից, ինչպես նաև օպերացիոն համակարգերից կամ լոգ ֆայլերից: Այս փուլում տվյալները հիմնականում վերցվում են իրենց սկզբնական, raw վիճակում:

Երկրորդ փուլը՝ Transform, որտեղ տվյալները վերափոխվում են միջանկյալ պրոցեսավորման միջավայրերում, օրինակ՝ ETL սերվերի վրա կամ հատուկ ETL գործիքի միջոցով: Վերափոխումը ընդգրկում է տվյալների մաքրում, ֆորմատների փոխակերպում, աղյուսակների միավորում, ագրեգացիաների և հաշվարկների ընդգրկում:

Վերջնին փուլը՝ Load, որը ապահովում է վերափոխված տվյալների պահպանումը նպատակային տվյալների պահեստում, սովորաբար Data Warehouse-ում, ինչպիսիք են Azure SQL Data Warehouse կամ Teradata: ETL ինտեգրման եղանակը արդյունավետ է, երբ անհրաժեշտ են ծանր և բարդ հաշվարկներ վերափոխման փուլում, և երբ նպատակային պահեստը չունի մեծ հաշվարկային հզորություն, երբ անհրաժեշտ է նախապես մաքրել և պահել միայն վավերացված տվյալներ, ինչպես նաև երբ տվյալները անկանոն կառուցվածքով են:

ELT-ն (Extract - Load - Transform) մոտեցումը տարբերվում է ETL-ից նրանով, որ գործընթացի փուլերի հերթականությունը փոխված է, և վերափոխումները կատարվում են հենց նպատակային տվյալների պահեստում: Սա դարձել է հնարավոր ժամանակակից ամպային տվյալների պահեստների մեծ հաշվարկային հզորությունների շնորհիվ: ELT-ում առաջին փուլը Extract է, որի ընթացքում աղբյուրներից ստացվում են տվյալները՝ առանց մեծ վերափոխումների: Երկրորդ փուլը՝ Load, ենթադրում է raw ֆորմատի տվյալների անմիջական պահպանումը տվյալների պահեստում, օրինակ՝ Azure Synapse Analytics, Snowflake, BigQuery կամ Databricks Delta Lake: Երրորդ փուլը՝ Transform, իրականացվում է հենց պահեստի ներսում՝ օգտագործելով SQL-ի, Spark-ի կամ MPP (Massively Parallel Processing) հաշվարկային մեխանիզմները:

ELT մոտեցման առավելությունն այն է, որ ժամանակակից պահեստները ունեն մեծ հաշվարկային հզորություն և կարող են արագ մշակել հսկայական ծավալի տվյալները: ELT-ը հատկապես օգտակար է, երբ աշխատում եք մեծ ծավալի տվյալների հետ, երբ պահեստը ունի բարձր հաշվարկային հզորություն, երբ վերափոխումները կարելի է իրականացնել SQL կամ Spark պահեստի ներսում, ինչպես նաև երբ անհրաժեշտ է արագ

բեռնել տվյալները և միայն հետո մշակել դրանք: Դրանց համեմատության աղյուսակը բերված է ներքևում:

Բնութագիր	ETL	ELT
Վերափոխման տեղը	Միջանկյալ համակարգ	Տվյալների պահեստում
Առավել հարմար է	Փոքր և միջին տվյալների համակարգերի համար	Մեծ տվյալների, ամպային պահեստների համար
Հաշվարկային ծանրաբեռնվածություն	ETL սերվերի վրա	Cloud/Data Warehouse compute-ի վրա
Տվյալների պահպանում	Պահպանվում են միայն վերափոխված տվյալները	Պահպանվում են raw ֆորմատի տվյալները, հետո վերափոխվում
Գործիքներ	SSIS, Informatica, Talend	Azure Synapse, Snowflake, BigQuery, Databricks
Սկզբնական ճկունություն	Ավելի կոշտ գործընթացներ	Շատ ճկուն՝ մոդելավորման և SQL transform-ի համար

OLTP/OLAP համակարգեր

OLTP (Online Transaction Processing) համակարգերը նախատեսված են օպերացիոն աշխատանքների համար, որտեղ առաջնահերթ են արագությունն ու կայունությունը, ինչպես նաև տվյալների մանրամասն վերահսկումը: Դրանք զբաղվում են առօրյա գործարքներով, որոնք փոքր ծավալով են, բայց տեղի են ունենում հաճախակի: Այս համակարգերը կարող են միաժամանակ սպասարկել հազարավոր կամ միլիոնավոր կարճ տևողությամբ տրանզակցիաներ՝ ապահովելով բարձր արագությունը և տվյալների ամբողջականությունը՝ պահպանելով ACID սկզբունքները, որոնցում ընդգրկված են Atomicity, Consistency, Isolation և Durability: Տվյալները սովորաբար պահվում են նորմալիզացված բազաներում՝ բազմաթիվ աղյուսակների բաժանված, որպեսզի նվազեցնեն կրկնօրինակումները և բարձրացնեն ամբողջականությունը: Օրինակների շարքում կարելի է նշել բանկային համակարգերը, որոնք ապահովում են հաշվիների միջև փոխանցումները և քարտային վճարումները, առցանց խանութները, որոնք իրականացնում են պատվերների գրանցում, վճարումներ և պահեստային փոփոխություններ, ինչպես նաև CRM համակարգերը, որոնք թույլ են տալիս օպերացիոն կառավարում հաճախորդների տվյալների:

OLAP (Online Analytical Processing) համակարգերը նախատեսված են վերլուծական և բիզնես-վերլուծական խնդիրների համար, երբ անհրաժեշտ է աշխատել մեծ ծավալի պատմական տվյալների հետ, կատարել բարդ հարցումներ և ստանալ խորքային վերլուծություն: Այս համակարգերը ապահովում են բարդ վերլուծական հարցումներ, որոնք կարող են ներառել խմբավորումներ, ագրեգացիաներ և բազմաչափ վերլուծություն: Տրանզակցիոն հաճախականությունը սովորաբար ցածր է, և հարցումները կարող են տևել վայրկյաններ կամ նույնիսկ րոպեներ: Տվյալները պահվում են դենորմալիզացված կառուցվածքով՝ օգտագործելով Star Schema կամ Snowflake Schema, ինչը ապահովում է վերլուծական հաշվարկների արագագործությունը: Այս համակարգերը հաճախ օգտագործվում են BI և Dashboard-ների ստեղծման համար, օրինակ՝ Power BI, Tableau և reporting համակարգերում: Օրինակներ են հանդիսանում Azure Synapse Analytics՝ ամպային մեծ տվյալների և վերլուծության հարթակը, ինչպես նաև Data Warehouse համակարգերը, ինչպիսիք են Snowflake, Google BigQuery և Amazon Redshift:

Տվյալային խողովակաշարեր (Data Pipelines)

Data Pipeline-ը տվյալների ավտոմատացված հոսքերի համակարգված հավաքածու է, որի միջոցով տվյալները վերցվում են տարբեր աղբյուրներից, մաքրվում, ձևափոխվում, մշակում և տեղափոխվում են նպատակային համակարգեր: Այն ապահովում է, որ տվյալները հոսեն շարունակաբար, կանխատեսելի և հուսալի ձևով՝ առանց մարդկային միջամտության: Data Pipeline-ը կարող է ներառել ինչպես պարզ գործողություններ, օրինակ՝ ֆայլի պատճենում, տեղափոխում, այնպես էլ բարդ գործընթացներ, ինչպիսիք են մեծ ծավալների վերլուծությունը, real-time տվյալների մշակումն կամ մեքենայական ուսուցման մոդելների թարմացումը:

Տվյալային խողովակաշարի հիմնական քայլերը սկսվում են տվյալների հավաքագրումով, որի ընթացքում տվյալները ստացվում են տարբեր աղբյուրներից, ինչպիսիք են տվյալների բազաները (SQL/NoSQL), API-ներն ու web services-ը, ֆայլային համակարգերը (CSV, JSON, Parquet) և IoT սարքերը: Այս փուլում տվյալները հիմնականում հավաքվում են իրենց raw տեսքով:

Մյուս փուլը ձևափոխումն է (ETL/ELT), որտեղ տվյալները անցնում են մաքրման, սխալների ուղղման, ֆորմատների ստանդարտացման, ֆիլտրավորման կամ բարդ վերափոխումների միջով: ETL-ի դեպքում վերափոխումները կատարվում են միջանկյալ համակարգում, իսկ ELT-ի դեպքում դրանք իրականացվում են տվյալների պահեստում՝ օգտագործելով SQL կամ Spark:

Վերջնական փուլը տվյալների տեղափոխումն է տարբեր պահեստներ. մշակված տվյալները տեղափոխվում են Data Warehouse՝ վերլուծության համար, Data Lake՝ մեծ չմշակված տվյալների պահպանման համար, կամ Operational Store և analytics engine: Տեղափոխումն կարող է կատարվել batch, near-real-time կամ real-time ձևաչափով:

Մոնիթորինգն ու սխալների կառավարումը ապահովում են, որ pipeline-ը աշխատի առանց ընդհատման: Այս փուլը ներառում է workflow-ի վերահսկում, սխալների ավտոմատ հայտնաբերում և վերագործարկում, ինչպես նաև լոգավորում և ծանուցումների հավաքագրում (alerting):

Վերջնական օգտագործման փուլում pipeline-ի միջոցով ստացված տվյալները օգտագործվում են վերլուծական հաշվետվությունների (BI dashboards) ստեղծման, մեքենայական ուսուցման մոդելների ուսուցման և թարմացման, օպերացիոն համակարգերի որոշումների կայացման, ինչպես նաև real-time համակարգերի ավտոմատացման համար:

Data Pipeline-ների կարևորությունն այն է, որ դրանք ապահովում են տվյալների հուսալի շրջանառություն՝ անկախ ծավալից կամ աղբյուրների բազմազանությունից, կրկնության դիմացկուն են՝ կարող են ինքնաշխատ վերագործարկվել կամ շարունակել սխալի դեպքում, ավտոմատացված են և նվազեցնում են ձեռքով միջամտության անհրաժեշտությունը, ինչպես նաև ապահովում են, որ տվյալները հասանելի լինեն ճիշտ պահին, ինչը կարևոր է վերլուծության, որոշումների կայացման և AI/ML համակարգերի համար: Pipeline-ների միջոցով համակարգերը դառնում են ճկուն, քանի որ տվյալները կարող են հոսել մեծածավալ և արագ աճող համակարգերի միջև:

Microsoft Azure Data Factory

Azure Data Factory-ն Microsoft Azure-ի ամպային տվյալների ինտեգրման ծառայությունն է, որը թույլ է տալիս կառուցել ամբողջական տվյալային հոսքեր՝ սկսած տվյալների հավաքագրումից մինչև դրանց մշակում, փոխակերպում և պահպանում տարբեր նպատակային համակարգեր: Այն հանդիսանում է որպես կապող օղակ տարբեր տվյալների աղբյուրների միջև՝ ապահովելով ավտոմատացված, վերահսկվող և մասշտաբավորվող տվյալային գործընթացներ: Azure Data Factory-ն լայնորեն կիրառվում է ETL և ELT գործընթացների իրականացման, Data Pipeline-ների ստեղծման ու կառավարման, բազմաթիվ աղբյուրներից տվյալների տեղափոխելը դեպի պահեստներ, ինչպես նաև տարբեր քայլերի հաջորդական կառավարման, ժամանակացույցի և սխալների մշակման համար: Ծառայությունը հնարավորություն է տալիս աշխատել ինչպես Azure-ի ներքին, այնպես էլ արտաքին կամ տեղային տվյալների աղբյուրների հետ, ապահովելով ճկուն ինտեգրում:

Azure Data Factory-ի հիմնական բաղադրիչներն են Pipelines, Activities, Linked Services, Datasets և Integration Runtime: Pipelines-ը ADF-ի հիմնական կառուցվածքային միավորն է, որը ներկայացնում է գործողությունների տրամաբանական հավաքածու՝ միաժամանակ կատարելով գործընթացներ տվյալների հետ, օրինակ՝ տվյալների ստացում, մաքրում և պահեստավորում: Pipeline-ները կարող են աշխատել ժամանակացույցով, արձագանքել տրիգերներին, պարունակել կախվածություններ և լինել մասշտաբավորվող ու վերահսկվող:

Activities-ը Pipeline-ի մեջ գործող քայլերն են, որտեղ ամեն activity կատարում է կոնկրետ առաջադրանք, ինչպիսիք են տվյալների պատճենումը, SQL սքրիպտի գործարկումը կամ API կանչը: Հիմնական activity-ներն են Copy Activity, որը իրականացնում է տվյալների փոխանցում աղբյուրից դեպի նպատակային պահեստ, Data Flow, որը ապահովում է տվյալների գրաֆիկական ձևափոխում Spark-ին նման parallel engines-ի վրա, Stored Procedure Activity, որը գործարկում է SQL stored procedure տվյալների բազայում, Web/HTTP Activity, որը կատարում է վեբ API-ների կանչ տվյալներ ստանալու կամ գործողություն անելու համար, և Execute Pipeline, որը թույլ է տալիս մեկ pipeline-ի կանչը մյուսից՝ բարդ workflows կառուցելու ժամանակ:

Linked Services-ը կապ է, որը սահմանում է, թե ինչպես է ստեղծվում կապը տվյալների աղբյուրի և նպատակային սերվիսի հետ: Այն պարունակում է connection string, authentication տվյալներ և endpoint հասցե: Օրինակներ են Azure Blob Storage, Azure SQL Database, On-premises SQL Server, Amazon S3, Salesforce և Oracle: Linked Services-ը ADF-ի օբյեկտների միջև կապերի հաստատման գործիքն է:

Dataset-ը ներկայացնում է այն տվյալային օբյեկտը, որի վրա աշխատում են activity-ները: Այն նկարագրում է տվյալների տեսակը, սխեման, ֆայլի ձևաչափը (CSV, JSON, Parquet): Dataset-ը տվյալներն ինքնին չեն, այլ դրանց ներկայացումն է:

Integration Runtime-ը ADF-ի հաշվարկային ենթակառուցվածքն է, որը իրականացնում է տվյալների փոխանցումն ու ձևափոխումը: Այն է engine-ը, որը ֆիզիկապես կատարում է Copy Activity-ը, Data Flow-ը և այլ transform-ներ: IR-ի

տեսակներն են Azure IR, որը ամբողջությամբ կառավարվում է Azure-ի կողմից և օգտագործվում է cloud-to-cloud տվյալների շարժի համար, աշխատելով առանձ սերվեր, Self-hosted IR, որը տեղակայվում է օգտագործողի սեփական on-premises կամ VM միջավայրում և թույլ է տալիս աշխատել firewalled միջավայրերում On-premises ↔ Cloud տվյալների շարժի համար, և Azure SSIS IR, որը նախատեսված է SSIS (SQL Server Integration Services) packages-ները փոխադրել Azure և աշխատում է որպես լիարժեք SSIS կատարման միջավայր ամպում:

Գլխի ամփոփում

Azure-ը ապահովում է միասնական միջավայր, որտեղ տվյալների հավաքագրումը, փոխանցումը, պահպանումը և վերլուծությունը իրականացվում են նույն հարթակի վրա՝ բարձր արդյունավետության և ավտոմատացման մակարդակով: Այս մոդելը թույլ է տալիս կազմակերպություններին արագ կառուցել տեխնոլոգիական լուծումներ՝ առանց բարդ ենթակառուցվածք կազմակերպելու:

Տվյալների շարժի և կառավարման միասնական մոտեցումը ապահովվում է Azure Data Factory-ի միջոցով, որը կարգավորում է տվյալների հոսքերի ամբողջական ցիկլը՝ թույլ տալով կապել տարբեր բիզնես համակարգեր, ավտոմատացնել ժամանակացույցով գործընթացներ և ապահովել տվյալների մշտական թարմացում: Սա դարձնում է տվյալային ճարտարապետությունը կանխատեսելի, վերահսկվող և ճկուն:

Azure-ի տվյալային ծառայությունները համակցված աշխատելով ապահովում են տվյալների ճշգրտությունը, վերահսկվող շարժ, բռնկումների և սխալների ավտոմատ մշակելիություն, ինչպես նաև տվյալների հասանելիություն տարբեր թիմերի և գործիքների համար:

Վերլուծության և որոշումների արագացման համար Azure-ի վերլուծական ծառայությունները անմիջապես օգտագործում են Data Factory-ի միջոցով բեռնված և մշակված տվյալները, ինչը թույլ է տալիս ստեղծել իրական ժամանակի dashboard-ներ, կատարել պատմական տվյալների խորքային վերլուծություն, արագ պատրաստել մեքենայական ուսուցման մոդելներ և ընդունել բիզնես որոշումներ:

Բիզնեսի պահանջներին համապատասխան Azure-ը ապահովում է ճկունություն՝ թույլ տալով տարբեր աշխատանքաներ տեղափոխել ամպ, ընդլայնել կամ նվազեցնել ռեսուրսները ըստ պահանջի: Սա հատկապես կարևոր է այն կազմակերպությունների համար, որոնք աշխատում են մեծ տվյալների ծավալներով կամ ունեն փոփոխական բեռնվածություն: