Allegato 2

Istituto Nazionale di Statistica

Concorso pubblico, per titoli ed esami a 18 posti aumentati a 24 a tempo indeterminato per il profilo di ricercatore di III Livello professionale bandito con Delibera Dop 919/2018 del 20/08/2018 e Dop 971/2018 del 17/09/2019

Area Data Science - Prima Prova d'Esame Scritta - Traccia n.1

La candidata o il candidato progetti un cruscotto informativo per l'analisi di dati relativi al parco circolante di auto sul territorio nazionale, alle emissioni inquinanti delle auto ed ai livelli di qualità dell'aria, sulla base delle seguenti specifiche.

Il sistema da realizzare è alimentato da tre sorgenti informative:

- *sorgente 1*: fornisce i dati su pratiche che vengono registrate al Pubblico Registro Automobilistico, limitatamente (per semplicità) alle seguenti tipologie: immatricolazione di un veicolo, passaggio di proprietà, cambio di residenza del proprietario, rottamazione del veicolo. Ogni pratica è caratterizzata dalla targa dell'auto, la classe di emissioni (Euro0, Euro1, ecc.), la cilindrata, la data della pratica, il tipo della pratica, il codice fiscale del proprietario del veicolo, e la sua città di residenza. Nel caso di un passaggio di proprietà, il codice fiscale si riferisce al nuovo proprietario del veicolo. Nel caso di un cambio di residenza, la città si riferisce alla nuova residenza. I dati sono organizzati in un'unica tabella relazionale.
- *sorgente 2*: fornisce dati sul territorio, che consideriamo limitati a città, numero di abitanti, provincia e regione. Questi dati sono in formato RDF. Si può assumere (per semplicità) che non ci siano variazioni nel tempo sul territorio.
- *sorgente 3*: fornisce dati giornalieri, in formato JSON, sulla qualità dell'aria, indicando per ciascuna città e per ciascuna data l'indice della qualità dell'aria, che è un valore intero compreso fra 0 e 500, da interpretare secondo la seguente classificazione: 0-50 qualità buona; 51-100 rischio moderato per la salute; 101-150 malsano per gruppi sensibili; 151-200 malsano; 201-300 molto malsano, 301-500 pericoloso.

Il cruscotto da realizzare deve offrire i seguenti servizi di consultazione e analisi dei dati:

- visualizzazione del valore della qualità dell'aria insieme al numero di veicoli circolanti, entrambi consultabili per città, provincia, regione, classe di emissione, cilindrata, giorno, mese, trimestre ed anno (si assuma che un veicolo sia circolante in una città ad una certa data se il proprietario del veicolo risiede nella città in quella data).
- analisi sull'andamento della qualità dell'aria nel tempo e sul territorio, con l'obiettivo di trovare associazioni e relazioni di correlazione interessanti tra l'indice della qualità dell'aria e le emissioni inquinanti dei veicoli, anche in ottica di previsione dell'andamento dell'indice di qualità sulla base di modifiche alla composizione del parco circolante di veicoli.

La candidata o il candidato discuta:

- le principali problematiche da affrontare relative all'acquisizione dei dati dalle sorgenti ed una loro possibile soluzione,
- una possibile architettura del sistema informativo che alimenta il cruscotto,
- la modalità di realizzazione dei servizi di consultazione ed analisi descritti in precedenza, indicando quale tipo di modello di analisi si intende applicare, l'impostazione dei parametri fondamentali del modello, le strategie per valutare la significatività dei risultati, ed eventualmente (ma non necessariamente) ulteriori tipologie di dati che si reputi opportuno utilizzare.

Per tutto ciò che nella descrizione dei requisiti del cruscotto informativo da progettare risulti non completamente chiaro o dettagliato e per tutto ciò che non è specificato, compresi eventuali ulteriori tipologie di dati che si ritiene opportuno utilizzare, o aspetti relativi al dimensionamento dei database sorgente, la candidata o il candidato può formulare, esplicitandole e giustificandole, opportune ipotesi ed assunzioni e svolgere la prova anche sulla base di esse.

omissis

Allegato 3

Istituto Nazionale di Statistica

Concorso pubblico, per titoli ed esami a 18 posti aumentati a 24 a tempo indeterminato per il profilo di ricercatore di III Livello professionale bandito con Delibera Dop 919/2018 del 20/08/2018 e Dop 971/2018 del 17/09/2019

Area Data Science - Prima Prova d'Esame Scritta - Traccia n.2

La candidata o il candidato progetti un cruscotto informativo per l'analisi di dati relativi all'allocazione del personale di una società di consulenza che opera in Europa ed al costo di gestione delle sedi della società, sulla base delle seguenti specifiche.

Il sistema da realizzare è alimentato da tre sorgenti informative:

- *sorgente 1*: la sorgente è costituita da tre tabelle relazionali: la prima tabella, per ciascun dipendente, fornisce informazioni su matricola e ruolo (analista, manager, ecc.), data e sede di assunzione, città della sede, data di cessazione del rapporto di lavoro (valorizzata solo in caso di rapporto di lavoro concluso); la seconda fornisce informazioni sulle assegnazioni a progetti dei dipendenti, indicando per ciascuna assegnazione la matricola del dipendente, il progetto a cui è assegnato e la data da cui il dipendente inizia a lavorare al progetto; la terza tabella contiene, per ciascun progetto, il suo codice identificativo, la sede in cui si svolge, la data di inizio e la data di fine del progetto, il numero stimato di dipendenti che servono nel progetto, nei diversi ruoli.
- *sorgente 2*: fornisce dati sul territorio, che consideriamo limitati a città, regione e nazione. Questi dati sono in formato RDF. Si può assumere (per semplicità) che non ci siano variazioni nel tempo sul territorio.
- *sorgente 3*: fornisce dati mensili, in formato JSON, sul costo di gestione delle varie sedi, indicando per ciascuna sede e per ogni mese il costo di gestione (che include costi variabili, quali, ad esempio, i costi per l'elettricità, per il materiale di cancelleria, per la mensa, ecc.).

Il cruscotto da realizzare deve offrire i seguenti servizi di consultazione e analisi dei dati:

- visualizzazione del numero dei dipendenti, effettivo e previsto in base alle stime sui progetti, insieme al costo di gestione di sede, in modo che siano consultabili per sede, città, regione, nazione, ruolo dei dipendenti, progetto, mese, trimestre ed anno. Si noti che un dipendente ad una certa data lavora in una sede se non ha cessato in precedenza il suo rapporto di lavoro con la società ed è assegnato ad un progetto attivo in quella data presso quella sede, oppure se non è assegnato a progetti in quella data e la sede coincide con quella di assunzione.
- analisi sull'andamento dei costi di gestione delle sedi nel tempo e sul territorio, con l'obiettivo di trovare associazioni e relazioni di correlazione interessanti tra i dati gestiti, anche nell'ottica di ottimizzare la distribuzione del personale per diminuire i costi di gestione.

La candidata o il candidato discuta

- le principali problematiche da affrontare relative all'acquisizione dei dati dalle sorgenti ed una loro possibile soluzione,
- una possibile architettura del sistema informativo che alimenta il cruscotto,
- la modalità di realizzazione dei servizi di consultazione ed analisi descritti in precedenza, indicando quale tipo di modello di analisi si intende applicare, l'impostazione dei parametri fondamentali del modello, le strategie per valutare la significatività dei risultati, ed eventualmente (ma non necessariamente) ulteriori tipologie di dati che si reputi opportuno utilizzare.

Per tutto ciò che nella descrizione dei requisiti del cruscotto informativo da progettare risulti non completamente chiaro o dettagliato e per tutto ciò che non è specificato, compresi eventuali ulteriori tipologie di dati che si ritiene opportuno utilizzare, o aspetti relativi al dimensionamento dei database sorgente, la candidata o il candidato può formulare, esplicitandole e giustificandole, opportune ipotesi ed assunzioni e svolgere la prova anche sulla base di esse.

omissis

Allegato 4

Istituto Nazionale di Statistica

Concorso pubblico, per titoli ed esami a 18 posti aumentati a 24 a tempo indeterminato per il profilo di ricercatore di III Livello professionale bandito con Delibera Dop 919/2018 del 20/08/2018 e Dop 971/2018 del 17/09/2019

Area Data Science - Prima Prova d'Esame Scritta - Traccia n.3

La candidata o il candidato progetti un cruscotto informativo per l'analisi di dati relativi ai veicoli utilizzati per il trasporto merci da una società di logistica che opera sul territorio nazionale, ed ai costi di gestione degli hub della società, sulla base delle seguenti specifiche.

Il sistema da realizzare è alimentato da tre sorgenti informative:

- *sorgente 1*: la sorgente è costituita da quattro tabelle relazionali: la prima tabella, per ciascun veicolo, fornisce informazioni sulla targa, la capacità di carico (un valore che indica il massimo carico trasportabile), la data in cui è stato acquistato, l'hub a cui è inizialmente assegnato, la data in cui il veicolo è stato dismesso (perchè venduto o rottamato), valorizzata solo in caso di dismissione; la seconda tabella indica per ciascun hub, la città in cui è situato; la terza tabella fornisce informazioni sugli arrivi dei veicoli presso gli hub, indicando per ciascun arrivo la targa del veicolo, l'hub in cui è arrivato e la data dell'arrivo; la quarta tabella fornisce informazioni sulle partenze dei veicoli dagli hub, indicando per ciascuna partenza la targa del veicolo, l'hub da cui è partito e la data della partenza. Si assuma, per semplicità, che un veicolo in una certa data possa arrivare in un solo hub oppure possa partire da un solo hub.
- *sorgente 2*: fornisce dati sul territorio, che consideriamo limitati a città, provincia e regione. Questi dati sono in formato JSON. Si può assumere (per semplicità) che non ci siano variazioni nel tempo sul territorio.
- *sorgente 3*: fornisce dati giornalieri, in formato XML, sul costo di gestione dei vari hub, indicando per ciascun hub e per ogni data il costo di gestione (che include costi variabili, quali, ad esempio, i costi per il rifornimento di carburante ed i costi di manutenzione per i veicoli che stazionano nell'hub in quella data).

Il cruscotto da realizzare deve offrire i seguenti servizi di consultazione e analisi dei dati

- visualizzazione del costo di gestione di un hub insieme al numero di veicoli che stazionano nell'hub ed alla capacità di carico che sviluppano, in modo che siano consultabili per hub, città, provincia, regione, giorno, mese, trimestre ed anno (si assuma che un veicolo stazioni in un hub in una certa data se il veicolo è arrivato in quell'hub in quella data oppure prima di quella data e non sia ancora ripartito).
- analisi sull'andamento dei costi di gestione degli hub nel tempo e sul territorio, con l'obiettivo di trovare associazioni e relazioni di correlazione interessanti tra i dati gestiti, anche nell'ottica di ottimizzare lo stazionamento dei veicoli per diminuire i costi di gestione.

La candidata o il candidato discuta

- le principali problematiche da affrontare relative all'acquisizione dei dati dalle sorgenti ed una loro possibile soluzione,
- una possibile architettura del sistema informativo che alimenta il cruscotto,
- la modalità di realizzazione dei servizi di consultazione ed analisi descritti in precedenza, indicando quale tipo di modello di analisi si intende applicare, l'impostazione dei parametri fondamentali del modello, le strategie per valutare la significatività dei risultati, ed eventualmente (ma non necessariamente) ulteriori tipologie di dati che si reputi opportuno utilizzare.

Per tutto ciò che nella descrizione dei requisiti del cruscotto informativo da progettare risulti non completamente chiaro o dettagliato e per tutto ciò che non è specificato, compresi eventuali ulteriori tipologie di dati che si ritiene opportuno utilizzare, o aspetti relativi al dimensionamento dei database sorgente, la candidata o il candidato può formulare, esplicitandole e giustificandole, opportune ipotesi ed assunzioni e svolgere la prova anche sulla base di esse.

omissis

Allegato 5

Istituto Nazionale di Statistica

Concorso pubblico, per titoli ed esami a 18 posti aumentati a 24 a tempo indeterminato per il profilo di ricercatore di III Livello professionale bandito con Delibera Dop 919/2018 del 20/08/2018 e Dop 971/2018 del 17/09/2019

Area Data Science - Seconda Prova d'Esame Scritta - Traccia n.1

Domanda 1:

La candidata o il candidato fornisca una definizione per la nozione di "modello concettuale dei dati", evidenzi gli scopi per i quali viene utilizzato, descriva le caratteristiche dei principali linguaggi di modellazione concettuale dei dati, e discuta anche la relazione che esiste fra questi e i linguaggi per la rappresentazione di ontologie in intelligenza artificiale.

Domanda 2:

Nella progettazione di un flusso di integrazione di dati provenienti da sorgenti informative differenti è necessario affrontare diversi problemi dovuti alle possibili eterogeneità esistenti fra le sorgenti. La candidata o il candidato illustri le fasi principali di un processo di integrazione finalizzato alla realizzazione di uno schema riconciliato dei dati, evidenziando i problemi caratteristici di ciascuna fase e le possibili soluzioni.

Domanda 3:

La candidata o il candidato illustri il concetto di Unsupervised Learning (apprendimento non-supervisionato), discuta le sue principali applicazioni e descriva un algoritmo per Unsupervised Learning di sua conoscenza.

Omissis

Allegato 6

Istituto Nazionale di Statistica

Concorso pubblico, per titoli ed esami a 18 posti aumentati a 24 a tempo indeterminato per il profilo di ricercatore di III Livello professionale bandito con Delibera Dop 919/2018 del 20/08/2018 e Dop 971/2018 del 17/09/2019

Area Data Science - Seconda Prova d'Esame Scritta - Traccia n.2

Domanda 1:

La candidata o il candidato elenchi i modelli di dati utilizzati nei sistemi di gestione dati NoSQL, indicando per ciascuno di essi le principali caratteristiche ed un ambito di applicazione in cui il modello stesso può risultare particolarmente adatto.

Domanda 2:

Con il termine Information Extraction (IE) si indica l'estrazione automatica di informazione strutturata da sorgenti dati non strutturate o semi-strutturate. La candidata o il candidato illustri i principali approcci all'IE di sua conoscenza, evidenziandone le più importanti caratteristiche, i vantaggi e gli svantaggi, con particolare riferimento all'estrazione di dati da documenti testuali.

Domanda 3:

La candidata o il candidato illustri il concetto di Supervised Learning (apprendimento supervisionato), discuta le sue principali applicazioni e descriva un modello per Supervised Learning di sua conoscenza.

omissis

Istituto Nazionale di Statistica

Allegato 2

Concorso pubblico, per titoli ed esami a 18 posti aumentati a 24 a tempo indeterminato per il profilo di ricercatore di III Livello professionale bandito con Delibera Dop 919/2018 del 20/08/2018 e Dop 971/2018 del 17/09/2019

Area Data Science - Seconda Prova d'Esame Scritta - Traccia n.3

Domanda 1:

La candidata o il candidato fornisca una definizione per il concetto di "Ontologia", nell'accezione comunemente intesa in ambito informatico, elenchi i principali linguaggi per la rappresentazione di ontologie, descrivendone le più importanti caratteristiche. Discuta inoltre discuta possibili utilizzi delle ontologie per la gestione dei dati nei sistemi informativi.

Domanda 2:

La candidata o il candidato illustri le principali caratteristiche di ciascuna fase del processo di Extraction-Transformation-Loading (ETL), tipicamente adottato per il popolamento di database centralizzati a supporto di processi decisionali e di data analytics.

Domanda 3:

La candidata o il candidato fornisca una classificazione delle principali tecniche di "data mining", fornendo per ciascuna di queste una descrizione generale, il tipo di informazioni implicite nei dati che può individuare, e la descrizione di uno scenario applicativo.

omissis